Anonimização de bases de dados empresariais

de acordo com a nova Regulamentação Europeia de Proteção de Dados

Frederico António Sá Oliveira Pinho
Mestrado em Segurança Informática
Departamento de Ciência de Computadores
2017

Orientador
Prof. Doutor Manuel Correia, DCC-FCUP

Co-Orientador
Prof. Doutor Luís Antunes, DCC-FCUP

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Agradecimentos

À minha cara-metade, por todo o apoio ao longo desta jornada

Resumo

Vivemos numa era digital.

O volume e a velocidade com que os dados necessitam de ser processados impõem a utilização de sistemas informáticos apoiados em bases de dados, em detrimento dos obsoletos arquivos físicos. Muitas dessas bases de dados constituem um ativo fundamental das empresas, através das quais são diariamente acedidas e armazenadas informações de colaboradores, clientes e fornecedores.

Parte da informação contida nesses sistemas empresariais corresponde, na prática, a dados pessoais; porém, por desconhecimento ou incúria, não são usuais as preocupações relacionadas com a recolha, encriptação ou anonimização de todos esses dados, confiando-se cegamente nos operadores e na fiabilidade dos processos de autenticação e autorização do *software*. Mesmo quando se tratam de dados históricos, sem aparente valor de negócio ou obrigatoriedade legal de salvaguarda, o risco associado à quebra de privacidade é usualmente ignorado e muita da informação acaba por ser mantida, de forma intacta, ao longo da vida útil da empresa.

O novo Regulamento Geral de Proteção de Dados vem criar uma disrupção com este *status-quo*. As empresas passam a ser responsabilizadas por quaisquer dados pessoais que tenham em sua posse, pelo que é incentivada a implementação do conceito "*Privacy by Design*", acautelando boas práticas logo na fase de implementação de sistemas e recolha de dados, e periodicamente com ações de atualização, eliminação, de-identificação, pseudonimização ou anonimização. As coimas são pesadas (podem ascender a 4% da faturação anual global ou a €20.000.000,00), pelo que muitas empresas veem uma oportunidade para renovar processos, sistemas e mentalidades.

Neste trabalho são descritos os desafios que as empresas terão de enfrentar a curto--prazo, em consequência deste novo quadro regulamentar. Para garantir a

minimização dos riscos de quebra de privacidade são apresentados diferentes modelos de anonimização, assim como um conjunto de ferramentas e técnicas que permitem a respetiva implementação, juntamente com os respetivos riscos e "*pitfalls*".

Por fim, em colaboração com empresas nacionais, é demonstrada a aplicação de técnicas de anonimização (*K-Anonymity, L-Diversity, Differential Privacy* e *Data Masking*) sobre quatro casos práticos reais:

- Anonimização de uma "Base de Dados de Candidatos associada a Processos de Recrutamento de Recursos Humanos", visando a realização de estudos de *marketing* e posicionamento de marca;
- De-identificação de uma "Base de Dados de Clientes", de forma a permitir a utilização dos dados em sistemas não produtivos, mediante risco controlado;
- Pseudonimização de uma "Base de Dados de Trabalhadores Externos" para utilização em ambientes de testes, respeitando a sintaxe, tipologias e dependências semânticas dos atributos;
- De-identificação de uma "Base de Dados de Colaboradores" com o objetivo de viabilizar a integração com outros sistemas da organização.

Constata-se que o desafio não reside apenas na anonimização dos dados, mas também na salvaguarda da qualidade da informação. Face ao inevitável *trade-off* entre privacidade e utilidade dos dados, é sugerido o envolvimento das áreas de negócio ao longo das sucessivas iterações do processo, apoiando a afinação dos modelos com vista a alcançar os objetivos traçados.

Palavras-Chave: privacidade, dados pessoais, anonimização, de-identificação, pseudonimização, risco de re-identificação, bases de dados empresariais, GDPR.

Abstract

We live in a digital era.

The volume and speed involved in data processing require elaborate computer systems based on databases, disregarding obsolete physical archives. Many of these databases are company fundamental assets, through which customer, employee and supplier information is accessed and stored on a daily basis.

Some of this information is actually personal data; however, due to lack of knowledge or negligence, companies are usually not concerned with personal data collection, encryption or anonymization, having blind faith in data operators and software authentication and authorization processes. Even when dealing with historical data, with apparently no business value or legal requirement, privacy risks are usually ignored, leaving most information intact through company lifetime.

This "*status-quo*" is being disrupted by the EU General Data Protection Regulation, forcing companies to be held responsible for any personal data they have in their possession. Consequently, the "Privacy by Design" concept is heavily promoted, enforcing standards at systems implementation and data collection phases, along with frequent actions to ensure data update, deletion, de-identification, pseudonymization or anonymization. Heavy penalties will be imposed (up to 4% of global annual turnover or € 20,000,000), taking many companies to seize this opportunity in order to renew processes, systems and mindset.

This paper describes some of the challenges that companies will face in the short term, as a consequence of this new regulatory framework. To aid in privacy risk minimization, different anonymization tools and techniques are proposed, along with several pitfalls and risk estimation.

Furthermore, it’s also presented the practical work carried out in collaboration with Portuguese companies, demonstrating how anonymity techniques (namely K-Anonymity, L-Diversity, Differential Privacy and Data Masking) can be applied on four real study-cases:

• Anonymization of a "Candidate Database” from recruitment processes of an HR Department, in order to carry out marketing studies and brand positioning;

• De-identification of a "Customer Database", allowing the use of data in non-productive systems, along with risk estimation measures;

• Pseudonymization of an "External Workers Database" for use in test environments, respecting the syntax, typologies and semantic dependencies of the attributes;

• De-identification of an "Employee Database" in order to allow integration with other corporate systems.

It became clear that the real challenge doesn’t lie only in anonymising data, but also in preserving information quality. Although there’s an inevitable trade-off between data privacy and utility, business areas should be involved throughout the successive iterations of the process (supporting the refinement of the models), in order to achieve the anonymization outlined objectives.

Keywords: privacy, personal data, anonymization, de-identification, pseudonymization, re-identification risk, corporate databases, GDPR.

"*Não procure uma falha, encontre uma solução*."

Henry Ford

Índice

Lista de Tabelas

Lista de Figuras

Lista de Abreviaturas

CE	Conselho da Europa
CNPD	Comissão Nacional de Proteção de Dados
CRM	Customer Relationship Manager
DEV	Ambiente de Desenvolvimento
DTAP	Development, Testing, Acceptance and Production
ERP	Enterprise Resource Manager
EUA	Estados Unidos da América
GDPR	General Data Protection Regulation (equivalente a RGPD)
GPU	Graphics Processing Unit
HIPAA	Health Insurance Portability and Accountability Act
ID	Código único identificador
IMDB	Internet Movie DataBase
NIF	Número de Identificação Fiscal
PIA	Privacy Impact Assessment
PII	Personal Identifiable Information
QLD	Ambiente de Qualidade
RGPD	Regulamento Geral de Proteção de Dados
RH	Recursos Humano
SaaS	Software as a Service
UE	União Europeia

1. Introdução

As sucessivas inovações tecnológicas do final do século XX, aliadas às melhores práticas de gestão que promovem a eficiência e eficácia dos serviços, levaram a que muitas empresas optassem por alicerçar o negócio em aplicações informáticas robustas. Sejam simples programas de faturação, contabilidade e salários, sejam complexos sistemas de gestão integrada (ERP, CRM, etc.), de uma forma geral assentam num princípio técnico indiscutível: a informação está armazenada em bases de dados relacionais.

Por outro lado, muitos arquivos físicos foram tornados obsoletos e alvo de processos de "transformação digital", convertendo caixas de arquivo (até então guardadas em caves com fechaduras de alta segurança) em dados e metadados, à distância de alguns cliques e uma palavra-passe introduzida numa aplicação informática.

Estes sistemas podem atingir uma importância tal que passam a ser parte integrante dos planos de continuidade do negócio ("*business continuity*"), obrigando ao desenho de elaborados processos de contingência com vista a assegurar a respetiva disponibilidade e integridade. Há, no entanto, uma terceira variável nesta equação que não pode ser menosprezada: a **privacidade dos dados pessoais**.

1.1 Identificação do problema

O novo Regulamento Geral de Proteção de Dados [1] impõe uma gestão rigorosa do tratamento de dados pessoais, implicando alterações nos processos e procedimentos em vigor nas organizações públicas e privadas.

Todas as operações que envolvam a recolha, consulta ou de alguma forma manipulação de dados pessoais no âmbito da atividade de uma empresa terão de estar sujeitas a um controlo apertado, assegurando que são preservados os direitos individuais dos cidadãos e que os riscos de quebra de privacidade são minimizados. Fomenta também que seja equacionada a eliminação dos dados pessoais "obsoletos"

que, embora presentes nas bases de dados da empresa, já não sirvam qualquer propósito de negócio.

Por outro lado, o Regulamento promove que, sempre que possível, os dados pessoais sejam de-identificados ou pseudonimizados, removendo, modificando ou substituindo as caraterísticas individuais por representações codificadas.

Em alternativa, a empresa poderá optar por anonimizar (de forma irreversível) uma parte desses dados, já que dessa forma deixam de ser considerados, à luz do Regulamento, dados pessoais. Esta não é, no entanto, uma operação isenta de riscos. Nos últimos anos têm vindo a público diversos casos de falhas na anonimização de base de dados pessoais [2,3], dos quais se pode inferir que o risco de re-identificação dos indivíduos está sempre presente, já que tal dependerá apenas dos recursos que um potencial atacante tenha ao seu dispor.

1.2 Objetivos e abordagem ao problema

Pretende-se com este trabalho demonstrar os riscos subjacentes às bases de dados empresariais, face à necessidade de proteção de dados pessoais e mitigação dos riscos de quebra de privacidade.

Foi efetuado um levantamento exaustivo da legislação aplicável à temática “Privacidade de Dados Pessoais”, permitindo assim enquadrar as implicações e responsabilidades que as empresas terão de acautelar com a entrada em vigor do novo Regulamento Geral de Proteção de Dados. Além de discutir os principais desafios, serão apresentadas as medidas a implementar a curto e médio-prazo pelas empresas, visando garantir a conformidade no tratamento e gestão dos dados pessoais.

Serão clarificados os conceitos de anonimização, de-identificação e pseudonimização (uma vez que não estão ainda normalizados na comunidade científica), sendo apresentadas as respetivas vantagens, riscos e limitações, em paralelo com a caraterização das principais técnicas de anonimização (nas vertentes de generalização, *randomization* e pseudonimização),

Foram pesquisados e analisados os casos mais paradigmáticos de falhas na anonimização de dados pessoais ocorridos nos últimos anos (tendo por base ataques cientificamente documentados). Será apresentado um resumo de cada um desses

ataques, assim como diversos cuidados a ter na aplicação de determinadas técnicas de anonimização, de forma a não incorrer em graves falhas de privacidade.

Por fim serão apresentados quatro casos práticos reais, com a demonstração de aplicação de modelos de anonimização, de-identificação e pseudonimização sobre bases de dados disponibilizadas por empresas nacionais. Em cada um dos casos foi fundamental a interação com os responsáveis e destinatários dos dados, de forma a permitir o desenvolvimento de modelos que respondessem às necessidades específicas de mitigação de riscos de quebra de privacidade, maximizando a preservação da qualidade e veracidade dos dados.

1.3 Organização do documento

Os próximos capítulos deste documento versam os seguintes temas:

- **Capítulo 2**: Enquadramento legal da proteção de dados pessoais a nível nacional e europeu, assim como uma análise das implicações do novo Regulamento Geral de Proteção de Dados no funcionamento das empresas;
- **Capítulo 3**: Conceitos de Anonimização, De-identificação, Pseudonimização, assim como diferentes técnicas e possível *software* de apoio;
- **Capítulo 4**: Falhas na implementação das técnicas de anonimização, potenciando elevados riscos de quebra de privacidade;
- **Capítulo 5**: Implementação prática sobre quatro casos reais de empresas portuguesas:
    - Anonimização de uma “Base de Dados de Candidatos associada a Processos de Recrutamento de Recursos Humanos”, visando a realização de estudos de marketing e posicionamento de marca;
    - De-identificação de uma “Base de Dados de Clientes”, de forma a permitir a utilização dos dados em sistemas não produtivos, mediante risco controlado;
    - Pseudonimização de uma “Base de Dados de Trabalhadores Externos” para utilização em ambientes de testes, respeitando a sintaxe, tipologias e dependências semânticas dos atributos;
    - De-identificação de uma “Base de Dados de Colaboradores” com o objetivo de viabilizar a integração com outros sistemas da organização.
- **Capítulo 6:** Conclusões e sugestões para trabalhos futuros.

2. Enquadramento legal

2.1 Dados pessoais

No direito da UE e do Conselho da Europa, «dados pessoais» são definidos como todas as informações relativas a uma pessoa singular identificada ou identificável, abrangendo aspetos respeitantes à vida privada, profissional ou pública; já nos EUA são usualmente designados por "PII" (*Personal Identifiable Information"*).

O conceito de "identificável" abrange todas as situações em que a pessoa, embora não explicitamente identificada, possa vir a sê-lo, caso sejam realizadas pesquisas cruzadas com outras fontes de dados.

É importante realçar que, mesmo encriptados, de-identificados ou pseudonimizados, por norma continuam a ser considerados dados pessoais.

A recente legislação preconiza também um regime especial para a salvaguarda de "dados pessoais sensíveis", incluindo dados de saúde, biométricos, genéticos, vida sexual, filiação sindical, origem étnica ou racial, assim como opiniões e convicções políticas e religiosas.

As bases de dados empresariais estão usualmente repletas de dados pessoais de:

- Colaboradores, incluindo:
    - Dados do Departamento de Recursos Humanos (recrutamento, seleção, vencimentos, formação, avaliação de desempenho, etc.);
    - Dados sensíveis do Departamento de Recursos Humanos (medicina do trabalho, fichas de aptidão médica, justificações de ausência por doença, tratamentos, gravidez, acidentes, etc.);
    - Extratos de telemóvel, portagens, cartões frota, etc.
- Fornecedores (incluindo dados dos respetivos trabalhadores);
- Clientes;

Por princípio, toda a informação guardada pela empresa no processo individual de um colaborador deverá constituir informação pessoal, já que permite inferir (direta ou indiretamente) um conjunto de dados pessoais. Também extratos de portagens, abastecimentos de combustível ou mesmo despesas de refeição constituem dados pessoais, já que permitem criar uma estreita ligação entre o colaborador e um objeto ou acontecimento.

O controlo dos dados pessoais pode tornar-se numa tarefa árdua numa empresa. Por exemplo, caso esteja implementado um controle ativo do ciclo de vida das aplicações informáticas no regime DTAP, é provável que todos os tipos de dados pessoais estejam a ser regularmente copiados para os ambientes de desenvolvimento e qualidade da empresa.

Em última análise, caberá a cada organização identificar todos os processos onde ocorra o acesso ou manipulação de dados pessoais, aferindo o risco e definindo procedimentos que assegurem o adequado tratamento destes dados, ponderando a hipótese de os anonimizar ou eliminar quando não puderem ser geridos [4], ou deixarem de servir a sua finalidade inicial.

2.2 Legislação – análise histórica

A primeira legislação europeia versando a proteção do direito à privacidade remonta a 1950, no rescaldo da 2ª Guerra Mundial, aquando da CEDH (Convenção Europeia dos Direitos do Homem): os art.º 8º, 9º e 10º estabeleceram a salvaguarda dos dados pessoais, em contraponto com o art.º 11º que preconiza a liberdade de acesso à informação.

O Código Civil Português, datado originalmente de novembro de 1966, passou a ordenar a tutela geral da personalidade (nas suas diferentes projeções), assim como o direito à reserva sobre a intimidade da vida privada (art.º 70º e 80º, respetivamente).

Por sua vez, logo na primeira versão da Constituição da República Portuguesa (abril de 1976) foi reconhecida a integridade moral e física dos cidadãos, mais tarde reforçada com o direito à reserva da intimidade da vida privada (art.º 26º) e com restrições à utilização da informática no âmbito do tratamento de dados pessoais (art.º 35º).

A Convenção 108 (1981) foi o primeiro instrumento internacional juridicamente vinculativo no domínio da proteção de dados, sendo aplicável a todos os tratamentos

de dados pessoais realizados tanto pelos sectores privado e público. A convenção permitiu que a proteção fosse alargada a pessoas coletivas (nomeadamente sociedades comerciais) mediante um contrato estabelecido entre as partes. Preconizou também a recolha e tratamento automatizado de dados de forma leal e lícita, armazenados para finalidades determinadas e legítimas, não autorizando a utilização para fins incompatíveis com essas finalidades nem conservados por tempo superior ao necessário. Dizia também respeito à qualidade dos dados, estabelecendo, em especial, que teriam de ser exatos, adequados, pertinentes e não excessivos (proporcionalidade).

A primeira diretiva europeia referente à Proteção de Dados - Diretiva 95/46/CE [5] almejou proteger as pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados, surgindo na sequência da necessidade de regulamentar a livre circulação de mercadorias, capitais, serviços e pessoas no espaço europeu. Esta diretiva foi transposta para a ordem jurídica portuguesa através da Lei n.º 67/98 (Lei da Proteção de Dados Pessoais). No entanto, verificou-se posteriormente que a aplicação não foi uniforme entre Estados-Membros, mantendo-se "*o sentimento generalizado na opinião pública de que subsistiam riscos significativos para a proteção das pessoas singulares, nomeadamente no respeitante às atividades por via eletrónica"* [1].

No ano 2000 foi finalmente assinada a Carta dos Direitos Fundamentais da União Europeia (a qual se tornou juridicamente vinculativa apenas com a assinatura do tratado de Lisboa de 2009) [6] e finalmente, em 2012, iniciaram-se os trabalhos para elaboração de uma proposta de revisão dos documentos existentes, a qual viria dar lugar ao Regulamento (UE) 2016/679 do Parlamento Europeu e o Conselho, usualmente conhecido como GDPR ("*General Data Protection Regulation*"), cuja tradução para a língua portuguesa estabeleceu a sigla RGPD ("Regulamento Geral de Proteção de Dados") [1].

2.3 Legislação – a nova regulamentação europeia

O novo Regulamento Geral de Proteção de Dados (GDPR) [1] foi publicado a 4 de Maio de 2016 no Jornal Oficial da União Europeia, estando prevista a sua total aplicação nos Estados-Membros a partir de 25 de Maio de 2018.

Uma vez que se trata de um Regulamento, as respetivas disposições são diretamente aplicáveis sem necessidade de qualquer transposição para cada jurisdição,

garantindo-se assim a verdadeira harmonização legislativa ao nível da Proteção de Dados em todos os países na União Europeia.

Mesmo que muitas empresas possam ter já adotado processos e procedimentos alinhados com a anterior Diretiva 95/46/CE [5], o novo regulamento impõe a rigorosa aplicação de boas práticas de segurança e privacidade dos dados (incluindo consentimento informado, pseudonimização, encriptação ou anonimização de dados, notificação de violação de privacidade, nomeação de "responsáveis" por dados pessoais nas empresas, etc.), obrigando à aplicação de reformas significativas nos processos e sistemas organizacionais, sob pena de pesadas coimas e penalizações.

Outra novidade oriunda deste Regulamento é a introdução do princípio "*Privacy by design and by default"* no tratamento dos dados pessoais, em paralelo com o incentivo à pseudonimização dos dados.

Um dos principais conceitos a ter presente durante a análise deste regulamento é o "tratamento de dados pessoais", definido como "*uma operação ou um conjunto de operações efetuadas sobre dados pessoais ou sobre conjuntos de dados pessoais, por meios automatizados ou não automatizados, tais como a recolha, o registo, a organização, a estruturação, a conservação, a adaptação ou alteração, a recuperação, a consulta, a utilização, a divulgação por transmissão, difusão ou qualquer outra forma de disponibilização, a comparação ou interconexão, a limitação, o apagamento ou a destruição".*

São também de destacar os seguintes considerandos da nova regulamentação, onde é apresentado o enquadramento da anonimização e pseudonimização aquando do tratamento de dados pessoais:

> Considerando 26: "*Os princípios da proteção de dados não deverão, pois, aplicar-se às informações anónimas, ou seja, às informações que não digam respeito a uma pessoa singular identificada ou identificável nem a dados pessoais tornados de tal modo anónimos que o seu titular não seja ou já não possa ser identificado. O presente regulamento não diz, por isso, respeito ao tratamento dessas informações anónimas, inclusive para fins estatísticos ou de investigação."*

> Considerando 28*: "A aplicação da pseudonimização aos dados pessoais pode reduzir os riscos para os titulares de dados em questão e ajudar os responsáveis pelo tratamento e os seus subcontratantes a cumprir as suas obrigações de proteção de dados. A introdução explícita da «pseudonimização» no presente regulamento não se destina a excluir eventuais outras medidas de proteção de dados."*

Tratando-se de uma substancial mudança de paradigma caberá às empresas reagir atempadamente, adaptando os procedimentos internos com vista a garantir a conformidade com o novo Regulamento. Um exemplo claro desta necessidade de antecipação é a subcontratação de pessoas e serviços que possam vir a lidar com dados pessoais: para tal, a redação dos Cadernos de Encargos deverá desde já incluir requisitos específicos sobre Proteção de Dados, tendo em conta as preocupações do novo Regulamento.

Os temas mais "sensíveis" e que deverão ser acautelados desde já pelas empresas encontram-se resumidos nas Tabela 1, 2 e 3:

GDPR: Medidas a implementar pelas empresas até maio/2018 **>** Componente **técnica / sistemas de informação:**	
art.12º **art.13º** **art.14º**	Assegurar a transparência nos procedimentos, facultando ao titular de dados pessoais todas as informações durante o processo de recolha, de forma a validar o respetivo consentimento (vd. 2.4);
art.32º	Aplicar pseudonimização e cifragem com vista a assegurar um nível de segurança (confidencialidade, integridade, disponibilidade e resiliência) adequado ao risco;
art.30º	Implementar mecanismos de auditoria e registo de todas as atividades de tratamento (incluindo consulta e modificação) de dados pessoais sob responsabilidade da empresa;
art.5.º **art.25.º** **art.47.º**	Implementar e comprovar processos de limitação, minimização, "*privacy by design"* e *"privacy by default*", reduzindo os dados tratados ao estritamente necessário, assim como a longevidade de conservação. Assegurar que os dados pessoais não são passíveis de ser disponibilizados a um número indeterminado de pessoas sem que exista uma prévia e controlada intervenção humana.
art.5º	Garantir a segurança dos dados, protegendo-os contra tratamentos não autorizados ou ilícitos, assim como perda, destruição ou danificação acidental.

Tabela 1 - GDPR – Medidas a implementar pelas empresas até maio/2018 – Componente técnica / sistemas de informação

GDPR: medidas a implementar pelas empresas até maio/2018 **>** Componente **organizativa/processual:**	
art.47º	Dar formação aos colaboradores que lidem com dados pessoais; Realizar frequentes auditorias sobre proteção de dados;
art.5º	Adotar todas as medidas adequadas para maximizar a qualidade dos dados, anulando ou retificando (sem demora) todos os dados inexatos;
art.28º + consid. 78º	Assegurar que as pessoas autorizadas a tratar os dados pessoais assumiram um compromisso de confidencialidade, e que os responsáveis pelo tratamento estão em condições de cumprir as suas obrigações em matéria de proteção de dados;

Tabela 2 - GDPR – Medidas a implementar pelas empresas até maio/2018 – Componente organizativa / processual

GDPR: medidas a implementar em "*ongoing*" pelas empresas:	
art.47º	Permitir que os titulares dos dados pessoais se oponham ou limitem o respetivo tratamento
art.35º **art.36º**	Avaliação de impacto sobre a proteção de dados. Solicitar um parecer ao "encarregado da proteção de dados" ou à CNPD sempre que o tratamento de determinados dados pessoais possa implicar um elevado risco;
art.33º **art.34º**	Notificar a CNPD e os titulares sempre que ocorra uma violação dos respetivos dados pessoais;

Tabela 3 - GDPR – Medidas a implementar em "ongoing" pelas empresas

São de realçar as categorias especiais de dados pessoais, com destaque para aqueles que explicitem (direta ou indiretamente) informações sobre a saúde e biometria de um indivíduo. Embora a regulamentação europeia permita o tratamento destes dados sensíveis no seio de uma empresa (com vista ao cumprimento de obrigações laborais e/ou contratuais), o art.9º exige que o tratamento ocorra "*sob a responsabilidade de um profissional sujeito à obrigação de sigilo profissional (…) ou confidencialidade ao abrigo do direito da União*".

As regras são aplicáveis não apenas à empresa responsável pelos dados, como também a todas as entidades que façam parte do grupo empresarial. Adicionalmente, todas as entidades em regime de subcontratação com acesso a dados pessoais do grupo empresarial terão, na sua essência, as mesmas obrigações que os responsáveis da empresa, pelo que estarão igualmente obrigados a cumprir o regulamento.

Por último, a empresa deverá documentar detalhadamente todas as atividades relacionadas com o tratamento de dados pessoais, de forma a conseguir comprovar a implementação das componentes técnicas e organizativas acima indicadas, assim como os resultados das frequentes avaliações de impacto sobre a proteção de dados (também designados por "PIA = *Privacy Impact Assessments*") (vd. 3.5 Riscos).

2.4 Recolha de dados e consentimento informado

É fundamental que as empresas garantam um correto processo de recolha de dados junto dos respetivos titulares. Para tal terão de ser prestadas todas as informações legalmente previstas de forma "*concisa, inteligível e de fácil acesso, utilizando uma linguagem clara e simples*" [7].

Os requisitos do processo de recolha encontram-se bastante detalhados no Regulamento Geral sobre a Proteção de Dados [1]. Assim, e de acordo com o artigo

6.º, por regra apenas será lícito o tratamento de dados pessoais por uma empresa quando:

- *"o titular dos dados tiver dado o seu consentimento para o tratamento dos seus dados pessoais, para uma ou mais finalidades específicas";*

ou

- *"o tratamento for necessário para a execução de um contrato no qual o titular dos dados é parte, ou para diligências pré-contratuais a pedido do titular dos dados.*

De acordo com o artigo 9.º do mesmo Regulamento, aquando da recolha, a empresa está obrigada, de uma forma geral, a prestar as seguintes informações aos titulares dos dados pessoais:

- Identidade e contactos do responsável pelo tratamento, incluindo os contactos do "Encarregado da Proteção de Dados" (quando aplicável);
- Finalidades do tratamento a dar aos dados pessoais, incluindo referências a eventuais interesses legítimos do titular dos dados;
- Destinatários dos Dados Pessoais (incluindo eventuais países fora da UE);
- Prazo de conservação dos dados recolhidos;
- Qual a obrigatoriedade de fornecimento dos dados (e.g. legal ou contratual), e eventuais consequências caso os dados não sejam fornecidos;

Quando a recolha for realizada por via eletrónica a empresa deverá registar e armazenar a evidência de um "consentimento informado". Este consentimento terá sempre de implicar uma ação intencional por parte do titular dos dados (i.e., não se poderá traduzir numa *checkbox* selecionada automaticamente). Por outro lado, o consentimento não é válido se "*a execução de um contrato, incluindo a prestação de um serviço, depender do consentimento, apesar de o consentimento não ser necessário para a mesma execução*" [1].

Caso alguma das informações prestadas venha a sofrer alterações, nomeadamente por se pretender realizar um tratamento distinto do inicialmente indicado, caberá à empresa a responsabilidade de informar novamente todos os titulares dos dados pessoais recolhidos.

Aos titulares dos dados pessoais terá de ser sempre permitido o direito de acesso, retificação, oposição, limitação do tratamento, portabilidade dos dados ou mesmo, em último caso, apagamento dos dados (“direito a ser esquecido”).

De realçar que, mesmo que tenham sido recolhidos no passado, cabe às empresas assegurar que os dados pessoais que têm em sua posse estão cobertos por um “consentimento do titular”, ou por outro fundamento legal legítimo, que cubra todas as possíveis utilizações dos dados. Tal poderá implicar a realização de contratos e/ou adendas em que o “titular dos dados” seja envolvido, com vista à regularização da situação.

2.5 O papel da CNPD

O art.º 51 do GDPR [1] designa que todos os Estados-Membros deverão estabelecer “autoridades de controlo”, responsáveis pela fiscalização da aplicação do Regulamento. Estas autoridades deverão ser entidades públicas, com total independência no exercício de poderes, diretamente nomeadas pelo Parlamento, Governo ou organismo equiparado.

No caso do Estado Português foi dada continuidade à atribuição realizada no âmbito da Lei 2/94 e da Diretiva 95/46/CE [5], nomeando-se a “Comissão Nacional de Proteção de Dados” (adiante designada por CNPD) como “Autoridade Nacional de Controlo de Dados Pessoais”.

Face à anterior legislação, o novo regulamento preconiza uma alteração de conceito no que diz respeito à interação entre as empresas e a CNPD. Foram, em grande parte, suprimidas as notificações obrigatórias (e respetiva emissão de autorizações) que consumiam significativos recursos da CNPD, dando lugar a uma maior responsabilização individual das empresas. Por outro lado caberá à CNPD apostar na sensibilização, prevenção, inspeção e repressão de ilícitos.

Competirá às empresas (e em especial ao encarregado de proteção de dados) colaborar com a autoridade de controlo, disponibilizando as necessárias informações que demonstrem o cumprimento do Regulamento.

É também responsabilidade das empresas avaliar os riscos de quebra de privacidade inerentes aos tratamentos de dados, solicitando um parecer à CNPD sempre que esse risco seja considerado elevado.

Por fim, caberá às empresas implementar os adequados procedimentos (abrangendo eventuais empresas subcontratadas) de forma a gerir corretamente situações de violações de segurança, documentando todos os incidentes e notificando a CNPD quando se conclua existir um significativo risco para os direitos dos titulares dos dados pessoais [7].

3. Conceitos de anonimização

Ao longo deste documento será utilizado genericamente o termo “anonimização” como referência a um processo de eliminação ou modificação da informação pessoal existente numa base de dados, com o objetivo de dificultar ou impedir a identificação unívoca dos indivíduos. Em pormenor, esse processo poderá traduzir-se na prática numa de-identificação, anonimização, ou pseudonimização, consoante as técnicas utilizadas e os objetivos alcançados.

A definição e desambiguação destes três termos foi expressa por [8,9,10] como:

De-identificação: consiste na remoção ou ofuscação de toda a informação pessoal de uma base de dados, com o objetivo de impedir a identificação dos indivíduos. A de-identificação não é necessariamente um processo irreversível, podendo prever-se a existência de uma tabela de mapeamento que permita reverter o processo (ligando os registos originais aos registos de-identificados).

Além da supressão de todos os atributos identificadores, a de-identificação usualmente implica a modificação dos “*quasi-identifiers”*, por via de processos de generalização (e.g. modificando a escala de um atributo) ou por introdução de fatores de incerteza, tendo por base os valores originais.

Anonimização: é considerado um caso “forte” de de-identificação, através do qual se pretende tornar impraticável, ou mesmo impossível (utilizando todos os meios considerados razoáveis) a re-identificação (inclusive pelo próprio técnico que realizou a operação inicial). Ou seja, por princípio deverá ser um processo irreversível, análogo à destruição. De realçar que o âmbito desta definição é adaptável consoante o contexto tecnológico do momento: “*todos os meios considerados razoáveis*”, permitindo assim ponderar os recursos necessários, o custo e conhecimento necessários para realizar uma re-identificação.

Pseudonimização: é um processo que visa substituir todos os identificadores pessoais (e.g. nomes, moradas e NIF) por pseudónimos: palavras ou códigos gerados artificialmente, os quais poderão funcionar como representações mascaradas dos dados originais. Uma pseudonimização "forte" tem adicionalmente a preocupação de incidir sobre os atributos "*quasi-identifiers*" (e.g. data de nascimento), e que a atribuição de códigos seja realizada de forma aleatória e independente dos valores originais (embora possam eventualmente continuar relacionados entre si). Por norma, a pseudonimização mantém todos os atributos de uma base de dados relacional, permitindo salvaguardar a respetiva estrutura e sintaxe dos dados.

3.1 De-identificação e anonimização

No tratamento de dados pessoais na área da saúde é bastante comum utilizar-se o termo "de-identificação", estando este inclusivamente explicitado na metodologia *HIPAA Safe Harbor*, a qual define como *standard* a remoção de 18 atributos "sensíveis" (vd. Tabela 4) como uma das condições necessárias para que uma base de dados possa ser considerada de-identificada [11].

1.Nomes	**7.** SSN ou NIF	**13.** Identificad. de dispositivos
2. Códigos postais (até 3 díg.)	**8**. N.ºs de registo médico	**14**. Web URLs
3. Datas (exceto o ano)	**9**. N.ºs de apólice de saúde	**15**. IPs
4. N.ºs de telefone	**10**. N.ºs de conta	**16**. Identificad. biométricos
5. N.ºs de fax	**11**. N.ºs de certificado ou licença	**17**. Fotos com face completa
6. Endereços de email	**12**. Identificad. de veículos	**18**. Outros atributos identificad.

Tabela 4 - HIPAA Safe Harbour - Atributos sensíveis

Por motivos técnicos ou funcionais poderão ser preservados os IDs únicos das tabelas iniciais, garantindo a ligação entre os registos e os indivíduos, e dessa forma permitindo uma posterior re-identificação (usualmente sujeita a um processo de autorização, já que implica o acesso a uma tabela de mapeamento ou mecanismo similar).

No entanto, mesmo os dados alvo de um processo de de-identificação continuam, à luz do GDPR, a ser considerados dados pessoais. Ao abrigo do considerando 26, apenas os dados "anónimos" estão excluídos dos princípios da proteção de dados: tal significa que a de-identificação seria insuficiente, sendo necessário um processo de efetiva anonimização (i.e., onde a re-identificação seja impraticável ou impossível).

De acordo com [10] a eficácia de um processo de de-identificação poderá ser aferida tendo como referência a dificuldade em:

1. Re-identificação dos indivíduos (também designado por "*single-out"*):
   De notar que a aplicação de uma simples máscara sobre o nome ou ID dos indivíduos, ou mesmo uma encriptação simétrica, permite sempre que a entidade que realizou a operação volte a identificar os indivíduos (i.e., são processos reversíveis).

2. Ligação entre registos, dentro ou entre bases de dados:
   Nomeadamente a capacidade de conectar ou correlacionar dois ou mais registos de um mesmo indivíduo (violando também a 1ª regra), ou de um grupo (neste caso violando apenas esta 2ª regra).

3. Inferência de qualquer informação sobre um indivíduo:
   Ou seja, deduzir o valor de um atributo de um indivíduo, com uma significativa probabilidade de sucesso, a partir de outros atributos.

Os processos de anonimização e de-identificação são por norma morosos, implicando diversas afinações até se obter a desejada utilidade dos dados finais. A situação ideal implicaria maximizar simultaneamente a privacidade e a utilidade dos dados (Fig. 1) – o que, por norma, é impossível de alcançar: as dificuldades advêm não só das limitações dos modelos, mas também da necessidade de interligação dos registos entre bases de dados de sistemas distintos.

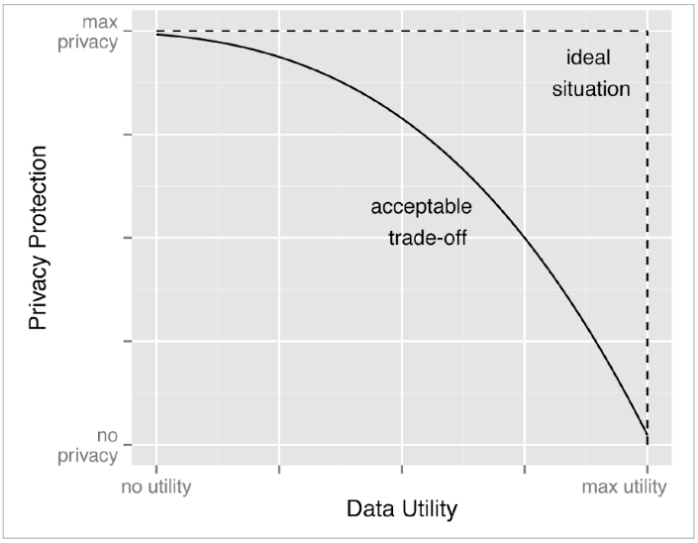


Fig. 1 - Anonimização: trade-off entre privacidade e utilidade dos dados [12]

Por outro lado, qualquer técnica de anonimização (ou mesmo a combinação de várias técnicas) possui inerente um risco de re-identificação dos indivíduos, nomeadamente pela combinação de *quasi-identifiers* (pedaços de informação suficientemente correlacionados, que potenciam a reconstrução de ligações entre os atributos e os indivíduos [10]). Daí ser crucial conhecer as caraterísticas e particularidades de cada técnica.

3.2 Pseudonimização

A perda de informação que resulta da aplicação de algoritmos de anonimização inviabiliza, muitas vezes, que os dados continuem a ser utilizáveis pelas empresas.

Além dos processos correntes de negócio (nos quais a informação tem de ser mantida sem perdas), as empresas poderão também ter necessidades de preservação de dados pessoais por requisitos de ordem legal (e.g. processos de despedimento).

Em termos técnicos esta preocupação também existe, nomeadamente quando estejam presentes complexos sistemas informáticos, com necessidades de comunicação e integração entre si, obrigando a que parte dos atributos de um indivíduo tenham que existir, sincronizadamente, em todos esses sistemas (algumas destas aplicações poderão estar na *cloud* suportadas em *SaaS,* assentando em diferentes tipologias de bases de dados).

Os cenários acima descritos dificultam a aplicação de técnicas de anonimização, já que obrigam a que pelo menos uma parte dos dados tenha que manter a qualidade e utilidade inicial.

Uma possível resposta a este problema passa pela utilização de pseudonimização, substituindo cada um dos atributos identificadores por um código que o permita mascarar.

Desta forma, caso se pretenda re-identificar o indivíduo associado a um conjunto de dados, será necessário consultar previamente uma tabela de mapeamento, armazenada separadamente. Essa tabela terá de estar sujeita a apertadas medidas de segurança, de forma a assegurar que não é acedida nem partilhada com entidades não autorizadas.

O processo deverá abranger não só os atributos identificadores de uma pessoa (usualmente conhecidos como “identificadores diretos”), mas também os restantes

atributos (potenciais "*quasi-identifiers*" – e.g. localidade de residência), salvo quando tal seja técnica ou funcionalmente inviável, nomeadamente quando exista a necessidade de manter a integridade de relacionamento desses atributos com outras bases de dados.

É sempre de realçar que a pseudonimização não é, por norma, suficiente para garantir, à luz da nova legislação europeia, que os resultados finais da operação deixem de constituir dados pessoais [13]. Adicionalmente, não pode ser negligenciado que a combinação dos dados pseudonimizados com outros conjuntos de dados poderá permitir a re-identificação total ou parcial dos indivíduos.

Mesmo assim, a pseudonimização constitui uma importante medida de segurança alinhada com o princípio de PbD (*Privacy by Design*) proposto na nova legislação. Funciona como uma técnica de substituição de dados pessoais, é relativamente simples de aplicar e permite reduzir a ligação, dentro do mesmo conjunto de dados, entre um indivíduo e os restantes atributos. Também a CNPD reconhece a viabilidade desta técnica sempre que não seja possível trabalhar com dados irreversivelmente anonimizados, desde que "*permita dificultar em grau elevado a identificabilidade dos sujeitos*" [14].

Outra das principais limitações da pseudonimização consiste na dificuldade em obter métricas que quantifiquem o efetivo risco de re-identificação, ou mesmo a taxa de sucesso do processo. Das diversas pesquisas efetuadas ao longo deste trabalho, apenas foram identificadas aplicações de similaridade sobre dados numéricos (e.g. [15]) ou limitados a categorias (e.g. [16]), existindo um aparente vazio no que diz respeito à aplicação, neste contexto, de algoritmos de similaridade de textos (e.g. distância de Levenshtein, Jaro-Winkler ou índice de Jaccard [17,18,19]). Como tal, a fiabilidade do processo estará em última análise dependente do analista responsável, não sendo possível garantir que foram realizadas corretamente todas as transformações necessárias para remover o caráter individual de cada registo.

3.3 Técnicas de anonimização

Ao longo deste documento será referido, de uma forma genérica, o termo "técnicas de anonimização", independentemente do objetivo pretendido: de-identificação, pseudonimização ou mesmo anonimização dos dados.

Embora a anonimização vise também manter a qualidade dos dados, por definição o processo implica inevitavelmente perda de informação, pelo que deverá ser acautelada a melhor relação entre:

- Utilidade dos dados: para permitir que os utilizadores consigam trabalhar com os dados, extraindo informação e estatísticas;
- Privacidade dos dados: para garantir que parte da informação se mantém ocultada.

De uma forma geral as técnicas de anonimização usualmente utilizadas podem ser categorizadas como *Randomization*, Generalização e Pseudonimização:

- ***Randomization***

Conjunto de técnicas que promovem a diluição da veracidade da informação, eliminando as fortes ligações entre os dados e cada indivíduo; baseiam-se na introdução de um fator de incerteza em determinados atributos, minimizando a associação entre os dados e cada um dos indivíduos.

É importante realçar que estas técnicas, isoladamente, não reduzem a singularidade de cada registo, embora permitam minimizar o sucesso de um eventual ataque por inferência.

De acordo com [10] as principais técnicas de anonimização por *randomization* são:

- ***Noise Addition***
  - consiste em aplicar ligeiras variações a valores numéricos e datas (e.g. adicionando +/- 2 meses a uma data);
  - permite ocultar os valores reais, embora seja difícil quantificar a variação necessária para minimizar os riscos de quebra de privacidade;

- ***Shuffling***
  - consiste em permutar, de forma aleatória, os valores de um atributo de uma mesma base de dados;
  - tem como pontos fortes o facto de permitir manter a tipologia e sintaxe dos dados originais;
  - como os dados finais são equivalentes aos iniciais (estando apenas organizados de forma diferente) poderá ser possível a reconstrução da organização inicial;

- ***Differential Privacy***
    - permite manter os registos de dados originais, adicionando ruído aleatório no resultado de cada pesquisa; o ruído gerado tem em conta as pesquisas realizadas previamente;
    - deverá obedecer ao princípio: "o resultado de um *query* não poderá contribuir para obter conhecimento sobre um indivíduo, estejam ou não os dados desse indivíduo na base de dados" [20];
    - possui como parâmetro principal ε (*epsilon*); quanto maior for ε, maior os riscos de quebra de privacidade [21];

- ***Generalização (Generalization)***

A generalização consiste na modificação da escala ou ordem de magnitude (e.g. substituindo o formato "mês/dia/ano" por apenas "ano").

As principais técnicas de anonimização por generalização são [10]:

- ***K-Anonymity***
    - consiste em agrupar K registos dentro de categorias ou intervalos de valores: grupo equivalente;
    - garante que existem 0 ou pelo menos K indíviduos (tuplos) que são abrangidos pelas mesmas combinações de *quasi-identifiers* (grupo equivalente); desta forma a probabilidade de identificação de um indivíduo é igual ou inferior a 1/K;
    - não impede que todos os indivíduos do grupo possuam o mesmo atributo sensível.

- ***L-Diversity / T-Closeness***
    - tratam-se de evoluções do método K-*Anonymity* com vista a garantir que cada classe de equivalência possui atributos suficientemente heterogéneos;
    - no caso da técnica *L-Diversity* terão de existir pelo menos L valores distintos por cada grupo equivalente e atributo "sensível";
    - já no caso da técnica *T-Closeness* procura-se aproximar a proporção de cada atributo "sensível" ao conjunto de dados originais, garantindo que T (*Threshold*) é a distância máxima entre as duas distribuições.

- ***Pseudonimização***

As técnicas de pseudonimização usualmente adotadas [10,22] para substituir ou mascarar dados pessoais são:

- **Substituição** (ou codificação)**:**
  - consiste em substituir um texto (e.g. nome do indivíduo) por um texto fixo, por um item obtido aleatoriamente a partir de uma lista de valores, etc.;
  - possui como pontos fortes a facilidade e simplicidade de aplicação, mesmo de forma não digital;
  - quando não aplicado corretamente poderá permitir facilmente a re-identificação dos indivíduos;

- **Encriptação** (usualmente com chave simétrica)
  - consiste na codificação dos dados usando uma chave pré-definida;
  - tratando-se de uma técnica determinística permite a re-identificação e recuperação dos dados originais;
  - por norma os dados codificados não respeitam a tipologia inicial;

- ***Hash***
  - consiste na codificação de dados unidireccionalmente, idealmente usando uma chave de entropia (*salt*) secreta;
  - apresenta como pontos fortes a facilidade e rapidez de aplicação;
  - por norma os dados codificados não respeitam a tipologia inicial;

- **Máscara de caracteres**
  - consiste em substituir caracteres de um texto por * (ou outro caractere pré-definido);
  - permite a aplicação em dados de texto, numéricos, etc.;
  - permite também ocultar apenas parcialmente um determinado texto ou atributo;

3.4 Análise comparativa

A Tabela 5 – adaptada de [10] - resume a eficácia das principais técnicas apresentadas no capítulo anterior:

	Risco de re-identificação dos indivíduos?	Risco de ligação entre registos?	Risco de Inferência de informação?
Randomization*: *Noise Addiction	Sim	Talvez	Talvez
Randomization: Shuffling	Sim	Sim	Talvez
Randomization: Differential Privacy	Talvez	Talvez	Talvez
Generalization: K-Anonymity	Não	Sim	Sim
Generalization: L-Diversity	Não	Sim	Talvez
Pseudonimização - Substituição	Sim	Sim	Sim
Pseudonimização - *Hash*	Sim	Sim	Talvez

Tabela 5 - Técnicas de Anonimização – Análise Comparativa

Na Tabela 6 – adaptada de [22] – são apresentadas as caraterísticas de diferentes técnicas de anonimização, facilitando o processo de seleção com base nos atributos que se pretendam tratar.

Técnica de Anonimização	Dados Reais	Mantida dimensão	Mantido formato	Determi-nístico	Difícil de reverter
Noise Addiction	Sim	Sim	Sim		Sim
Shuffling	Sim		Sim		
Substituição	Se necessário	Se necessário		Sim: tabela mapeamento	Sim, após elimin tabela
Encriptação				Sim	
Hash				Sim	Sim [1]
Máscara de caracteres		Sim		Sim	Sim

Tabela 6 - Técnicas de Anonimização – Principais Caraterísticas

De uma forma geral, a técnica de anonimização adequada dependerá sempre dos dados de origem e dos objetivos a alcançar. Poderão também ser obtidos bons resultados combinando mais que uma técnica no mesmo processo de anonimização.

Não obstante, sempre que for necessário manipular bases de dados relacionais, poderá ser necessário garantir que [22]:

- ➔ a anonimização preserva a estrutura física da base de dados;
- ➔ é reconstruída a integridade referencial entre as diferentes tabelas de uma ou mais bases de dados;
- ➔ são verificados os índices, chaves primárias e *triggers* da base dados, de forma a garantir que mantêm as caraterísticas da base de dados original.

[1] Irreversível por função direta; no entanto, caso a função *hash* e o *salt* sejam conhecidos, os dados poderão ser parcialmente recuperados por via de um ataque *brute-force* baseado em palavras de um dicionário;

3.5 Riscos

Qualquer processo de de-identificação detém intrinsecamente um fator de risco por quebra da privacidade, o qual é fortemente influenciado pelo contexto em que os dados são disponibilizados e pelas técnicas de anonimização adotadas. Por mais rigoroso que seja o processo de de-identificação, a re-identificação dos indivíduos poderá ser, em teoria, sempre possível. Basta para tal que sejam empregues recursos ilimitados, procurando correlações com dados ou indivíduos previamente identificados a partir de outras bases de dados.

É de realçar que este risco será tendencialmente superior sempre que o ataque vise um alvo específico (e.g. um único indivíduo), face à possibilidade de obter todo o tipo de informação externa, cruzando-a com os dados anonimizados.

Para quantificação do risco a que ficam submetidos os dados após serem sujeitos a processos de anonimização, algumas aplicações de *software* adotam modelos pré--definidos de estimativa de risco, com base nas probabilidades de uma re-identificação com sucesso. Ao longo deste documento será adotado o modelo dos três perfis de "*Prosecutor*", "*Journalist*" e "*Marketer*" [12], descritos como:

- *Prosecutor*. procura re-identificar um determinado registo, sendo assumido que possui já a confirmação da presença do indivíduo no conjunto de dados em causa;
- *Journalist*: procura re-identificar um determinado registo, embora sem confirmação se o indivíduo está presente no *subset* de dados analisado;
- *Marketer*: está interessado em re-identificar grandes volumes de registos, e não apenas registos individuais.

De realçar, no entanto, que os resultados apresentados nas estimativas de risco assentam na premissa que um potencial atacante terá apenas acesso ao conjunto de dados disponibilizado, o que muitas vezes poderá não ser verdade [23,24,3]. De forma a precaver uma falsa sensação de segurança, e uma vez que os temas de anonimização e re-identificação estão bastante ativos no meio científico, com frequentes publicações de pesquisas, descobertas e regulamentações, caberá às empresas assegurar que todos os riscos associados aos dados pessoais são frequentemente reavaliados.

A nova Regulamentação prevê explicitamente a necessidade das empresas realizarem avaliações de risco associado aos diversos tratamentos de dados pessoais, estando

obrigadas a comunicar à CNPD situações em que o risco resultante (não mitigável) seja considerado elevado.

É de realçar que, ao contrário dos habituais processos organizacionais de gestão de risco, a aceitação dos riscos residuais de privacidade terá de ser justificada e frequentemente reavaliada pela empresa, já que poderá constituir uma violação dos direitos dos cidadãos [25].

A avaliação de riscos poderá também ser complementada com a realização periódica de *Privacy Impact Assessments* (PIA), no qual são auditados os diversos processos organizacionais, com o objetivo de avaliar os riscos de quebra de privacidade subjacentes à atividade da empresa, assim como identificar eventuais ações mitigadoras [25]. Uma vez que a realização destas auditorias implica a avaliação do impacto das operações de tratamento, têm também a vantagem de promover implicitamente o cumprimento do Código de Conduta estipulado no art.º 40 do GDPR.

3.6 *Software* de anonimização

A utilização de técnicas de anonimização carece do apoio de *software* especializado, de forma a permitir a aplicação dos algoritmos de forma sistemática, com base em modelos previamente estabelecidos.

Num trabalho desenvolvido em 2014 por *Bergeat et al.* [26] foi listado e caraterizado o *software* habitualmente utilizado para aplicação de técnicas de anonimização, conforme apresentado na Tabela 7 (adaptado e atualizado de [26]).

Software	**Última versão**	**Promotor**	**Open Source**	**Funções de Anonimização disponibilizadas**
ARX	3.5.1 jan/2017	Munich University	Sim	*K-anonymity, L-diversity, T-closeness, δ-disclosure, δ-presence (ε, δ)-differential privacy Generalization, Suppression, Microaggregation, Top/bottom-coding, Global+Local recoding*
μ-Argus	5.1.1 abr/2015	Eurostat + Statistics Netherlands	Sim	*Global recoding, Local suppression, Top/bottom coding, Post Randomization, Additive noise, Microaggregation, Numerical Rank Swapping, Syntethic Data*

Software	Última versão	Promotor	Open Source	Funções de Anonimização disponibilizadas
sdcMicro	5.0.2 mai/2017	Vienna University of Technology	Sim	É um pacote R que disponibiliza funções de *Microaggregation, Additive noise, Data swapping, Local suppression, Partially synthetic data*
Privacy Analytics Eclipse	N/D	Privacy Analytics Inc.	Não	*FPE (format preserving encryption), pseudonymization, generalizations, date shifting, randomization, risk-based suppression, re-coding*
CAT	1.0 jun/2009	Cornell University	Sim	*K-anonymity,Incognito, L-diversity, Risk analyzer*
UTD	mar/2010	University of Texas (Dallas)	Sim	*K-anonymity,Datafly, Mondrian Multidimensional, Incognito, Incognito with L-diversity + Tcloseness*

Tabela 7 - Software de Anonimização

A maioria deste *software* tem origem em instituições académicas, motivo pelo qual são tendencialmente de código-livre (*open-source*).

Já para aplicação das técnicas de pseudonimização poderão ser utilizados simples *scripts* e *queries* SQL, ou, em alternativa (caso se pretenda um controlo acrescido sobre todo o processo), adotar soluções comerciais, sendo as mais comuns as seguintes:

- Oracle Data Masking
- IBM InfoSphere Optim Data Masking
- Informatica Data Masking Tool

Estes três produtos disponibilizam soluções de Dynamic Data Masking (online) ou Persistent Data Masking (em offline, visando manter a integridade referencial das diferentes bases de dados, assim como a propagação dos mesmos dados mascarados por diferentes aplicações).

No âmbito deste trabalho, em especial na aplicação prática aos casos de estudo *[vd. 5. Casos de estudo],* foi utilizado o seguinte *software*:

- ARX Anonymization v3.5.1 (para aplicação de técnicas de *randomization* e generalização);
- SQL + Excel scripting (para aplicação de técnicas de pseudonimização).

4. Falhas na anonimização

4.1 Introdução

Nos últimos anos têm vindo a público diversos casos reais demonstrando falhas em processos de anonimização. Estas situações têm evidenciado a necessidade do envolvimento de especialistas nos processos de anonimização, já que qualquer erro poderá gerar repercussões na imagem da organização. Com a entrada em vigor da nova Regulamentação, as penalizações poderão ser ainda maiores, já que o art.º 82º do GDPR preconiza que "*qualquer pessoa que tenha sofrido danos materiais ou imateriais devido a uma violação do presente regulamento tem direito a receber uma indemnização do responsável pelo tratamento ou do subcontratante pelos danos sofridos*" [1].

É importante realçar que a remoção de atributos identificadores (nome, telefone, etc.) nunca será suficiente para mitigar os riscos de quebra de privacidade. Por outro lado, deverá sempre assumir-se que um potencial atacante tem acesso a outras bases de dados externas, podendo cruzar essa informação com os dados que se pretendam anonimizar.

Nos capítulos seguintes serão apresentados exemplos de ataques que permitiram a re-identificação de indivíduos em bases de dados supostamente "anonimizadas", assim como um conjunto de falhas bastante frequentes em determinadas técnicas de anonimização, e respetivas sugestões de mitigação.

4.2 Exemplos de ataques de re-identificação

Foi realizado um trabalho de pesquisa nas plataformas *IEEE Xplore*, *Semantic Scholar* e *Google Scholar,* de forma a identificar exemplos paradigmáticos de re-identificação de dados pessoais, tendo os resultados sido filtrados com base na relevância e

categorizados conforme proposto por [8], nomeadamente: “de-identificação insuficiente”, “ligação a outra base de dados” e “reversão dos pseudónimos”:

4.2.1 De-identificação insuficiente

- **O caso AOL (2006)**

O grupo America OnLine (AOL) decidiu divulgar publicamente uma base de dados com 20 milhões de pesquisas realizadas por 650 mil clientes ao longo de três meses, com o objetivo de apoiar investigações académicas. Foram removidos da base de dados todos os atributos identificadores e substituído o nome do cliente por um código único aleatório. No entanto, jornalistas do New York Times [27,28] constataram que muitas das pesquisas explicitavam informações pessoais (nomeadamente o próprio nome e residência), pelo que, com base no código único, foi possível inferir todas as restantes pesquisas efetuadas por determinados indivíduos (incluindo dúvidas médicas e sexuais).

- **Registos de hospitalização do estado de Washington, EUA (2011)**

O Estado de Washington vendeu por $50 um conjunto de dados médicos contendo todas as hospitalizações ocorridas no ano de 2011; embora não explicitando dados identificadores, a informação incluía os diagnósticos, as intervenções realizadas, informações sobre o médico e hospital, assim como os custos da hospitalização. Na investigação realizada por [29] foi possível demonstrar a re-identificação exata e unívoca de 35 pacientes, tendo apenas por base as notícias publicadas em jornais durante esse ano.

- **Metadados de Telemóveis e Cartões de Crédito (2013,2015)**

O comportamento humano está expresso na forma como cada indivíduo atua no seu dia-a-dia, obedecendo a um conjunto de padrões tendencialmente repetitivos.

De acordo com o trabalho de investigação levado a cabo por *Montjoye et al*. [30,31], os registos de utilização do telemóvel ou cartão de crédito permitem identificar (*single--out*) univocamente a maioria dos seus utilizadores.

Analisando uma base de dados com 1.5 milhões de registos de chamadas telefónicas, a equipa de investigadores verificou que bastariam quatro observações aleatórias (local e hora) para re-identificar 95% dos indivíduos. Estas observações poderiam ser obtidas com base em diversos *inputs* do próprio indivíduo (e.g. redes sociais, IP de

origem no envio de emails, compras com cartão de crédito, etc.), e permitiriam inferir a respetiva localização ao longo de um período de 15 meses.

Tendo por base três meses de registos de utilização de cartões de crédito, de um total de 1.1 milhões de clientes, foi também possível demonstrar que bastariam quatro pontos espaço temporais aleatórios (dia e localização) para re-identificar univocamente 90% dos clientes, e a partir daí conhecer todo o seu historial de compras. Caso fosse também conhecido o valor aproximado da transação, bastariam apenas três pontos (dia, localização e valor) para alcançar o mesmo objetivo.

4.2.2 Ligação a outra fonte de dados

- **O caso NETFLIX Movie Ratings (2007)**

Em 2007 a organização NETFLIX divulgou publicamente uma base de dados contendo 100 milhões de avaliações realizadas por 500 mil clientes, substituindo os nomes por IDs aleatórios, tendo também sido adicionada alguma perturbação/variação nas avaliações.

Cruzando esta base de dados com a informação disponibilizada pela IMDB, investigadores da Universidade de Texas-Austin [32] demonstraram que 99% dos registos poderiam ser potencialmente re-identificados, bastando para tal encontrar oito filmes em comum nas duas bases de dados. Com este exercício foi possível explicitar preferências políticas e outras informações sensíveis de clientes previamente identificados.

Face à multidimensionalidade dos dados de avaliações do sistema NETFLIX, a aplicação de técnicas de anonimização (como *K-Anonymity* e *Differential Privacy*) dificilmente permitiria manter a utilidade dos dados, pelo que, face ao risco envolvido, a divulgação pública desta informação seria sempre de evitar.

- **Os registos médicos do Governador de Massachusetts (1997)**

A Massachusetts Group Insurance Commission (GIC) disponibilizou publicamente uma base de dados de informação médica “de-identificada” (incluindo data de nascimento, código postal, sexo, datas de hospitalização, diagnósticos, resultados dos exames, receitas e custos incorridos), com o objetivo de promover a melhoria dos cuidados de saúde e o controlo dos respetivos custos. A comissão, com o apoio do Governador

William Weld, assegurou que a privacidade dos doentes estava acautelada, uma vez que tinham sido eliminados os identificadores pessoais.

A investigadora Latanya Sweeney argumentou que uma grande parte dessa informação médica poderia ser facilmente re-identificada com base em apenas três *quasi-identifiers*: data de nascimento, sexo e código postal [33,3,2]. Para suportar esta afirmação cruzou a informação médica com a lista de eleitores da cidade de Cambridge – Massachusetts (adquirida por $20), re-identificando toda a informação médica do Governador, uma vez que apenas seis pessoas hospitalizadas partilhavam a mesma data de nascimento, das quais apenas três eram do sexo masculino; o último atributo permitiu completar a re-identificação, já que dos três potenciais candidatos apenas o Governador vivia num determinado código postal.

A ligação entre estas duas bases de dados poderia gerar falsos positivos (uma vez que não é garantido que todos os indivíduos estejam presentes em ambas), pelo que diversos autores [34,23,8] consideram que seria fundamental cruzar as "suspeitas de re-identificação" com uma terceira fonte de informação.

De notar que após 2003 a divulgação desta informação médica não seria permitida, uma vez que violaria o HIPAA.

- **Re-identificação de amostras de DNA (2013)**

À semelhança dos apelidos masculinos, também algumas propriedades do cromossoma Y são transmitidas, de forma inalterada, de pais para filhos. Yaniv Erlich demonstrou ser possível re-identificar cerca de 50 genomas masculinos (doados "anonimamente" para investigações científicas), através de correlações com bases de dados recreativas de genealogia, disponíveis publicamente na internet [35,36,9].

As bases de dados de investigação científica de genomas incluem diversos metadados, entre os quais a idade e residência (estado) de cada um dos dadores. Filtrando a pesquisa por indivíduos com caraterísticas genéticas similares (polimorfismos de nucleotídeo único), foi possível identificar os apelidos de potenciais candidatos em duas bases de dados *online* de genealogia (YSearch e SMGF). A re-identificação foi concluída complementando estas informações com pesquisas no *Google* e consulta de registos de obituários, conseguindo eliminar os restantes candidatos.

De referir que atualmente, a partir de 69€, qualquer cidadão português poderá solicitar online a sequenciação do seu DNA, utilizando serviços como o MyHeritage, EasyDNA,

ou AncestryDNA (empresas sediadas em Israel, UK e EUA). Seria expectável que, após elaborado o relatório, os dados fossem eliminados: no entanto, nenhum destes serviços apresenta garantias sobre a privacidade da informação genética.

4.2.3 Reversão dos Pseudónimos

- **O caso NYC Taxi (2014)**

Um exemplo de uma falha deste tipo ocorreu em 2014 em Nova Iorque, resultado da divulgação pública de dados (supostamente "anonimizados") sobre 173 milhões de viagens de táxi realizadas na cidade [37]. O registo de uma dessas viagens está representado na Fig. 2, onde constam dois atributos hexadecimais de 32 caracteres, resultantes da aplicação da função de *hash* MD5.

```
6B111958A39B24140C973B262EA9FEA5,D3B035A03C8A34DA17488129DA581EE7,VT
S,5,,2013-12-03 15:46:00,2013-12-03
16:47:00,1,3660,22.71,-73.813927,40.698135,-74.093307,40.829346
```

Fig. 2 - Ataque por reversão de *hash* - exemplo de um registo utilizando MD5

Estes dois atributos correspondem ao "código do táxi" e ao "nº de licença do taxista", ambos com sintaxe publicamente conhecida, sendo facilmente reversíveis usando um ataque por "*brute-force*". Neste caso bastariam cerca de 22 milhões de cálculos para obter todas as possíveis combinações (revelando para o registo acima apresentado os códigos 8T19 e 5092438), e consequentemente permitindo obter a totalidade da base de dados original.

4.3 *K-Anonymity* - Ataques por composição

Uma das principais falhas da técnica de *K-Anonymity* (incluindo as derivações *L-Diversity* e *T-Closeness*) reside na possibilidade de se efetuar um ataque por composição.

O exemplo clássico deste tipo de ataques está representado na Fig. 3 [38] e pode ser descrito como:

- Uma população é servida por dois hospitais, aos quais podem aceder livremente todos os cidadãos;

- Cada hospital possui uma base de dados com informação sobre as doenças de cada utilizador;
- Ambos os hospitais divulgam publicamente estatísticas (usando técnicas de *K-Anonymity* e *L-Diversity/T-Closeness*, limitando a identificação unívoca);
- Cruzando a informação disponível nas duas bases de dados passa a ser possível identificar univocamente os doentes, inferindo as respetivas doenças.

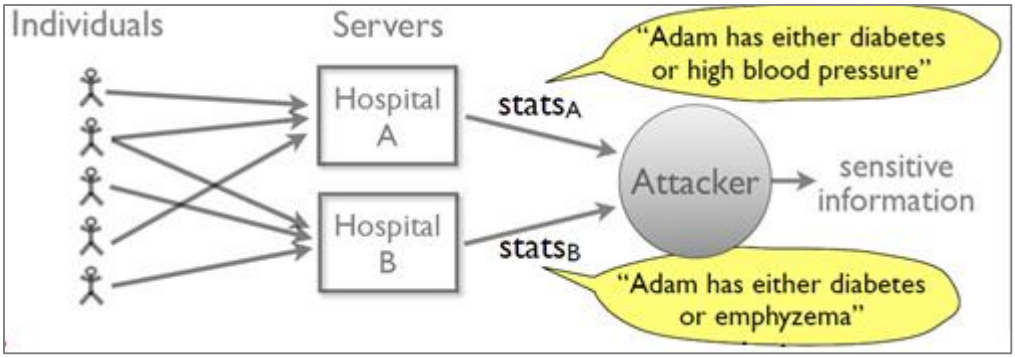


Fig. 3 - Ataque por composição de duas bases de dados "anonimizadas" por *K-Anonymity*

De acordo com [38] este tipo de ataques por composição pode ser mitigado usando técnicas de *Differential Privacy*.

4.4 *Hash* - Ataques por reversão

Muito embora uma função *hash* seja, por definição, unidirecional, tal não implica que esta seja tecnicamente irreversível. De facto, se a envolvente de dados base e o algoritmo *hash* forem publicamente conhecidos, um atacante poderá recalcular todos os valores possíveis com vista a determinar os dados originais.

À semelhança do indicado em *[4.2.3 Reversão dos Pseudónimos]*, também poderá ocorrer um ataque deste tipo sobre uma base de dados empresarial. Tome-se como exemplo o NIF (número de identificação fiscal): é composto por 9 dígitos, dos quais o último é um dígito de controlo, pelo que existem no máximo $1x10^8$ combinações possíveis. Caso fosse adotada uma codificação deste atributo usando o algoritmo de *hash* SHA1, e admitindo que a utilização de *GPUs* permite atingir taxas de 4000MH/s [39], facilmente se depreende que todas as combinações de NIF poderiam ser calculadas em menos de um segundo.

Em alternativa poderia ser adotado um algoritmo de *hash* utilizando um *salt* dinâmico, obtido por combinação entre uma chave confidencial e uma representação do conteúdo do próprio registo.

4.5 *Shuffling* - Ataques por dedução incremental da tabela de mapeamento

A técnica de *Shuffling* preconiza que a cada *string* original deverá corresponder uma e só uma *string* mascarada, permitindo assim que o processo de *masking* garanta a integridade e uniformidade em todas as tabelas ou bases de dados relacionadas.

Exceto quando possua relevância para o destinatário final, a fonte de dados mascarados deverá ser distinta da base de dados original. De outra forma um atacante terá inevitavelmente conhecimento sobre o universo de dados originais e, consequentemente, poderá confirmar se determinado registo consta ou não da base de dados. Por exemplo, seria simples confirmar se determinada empresa consta, ou não, de uma lista de fornecedores.

Por vezes os processos de *shuffling* decompõem as *strings* e atuam individualmente sobre cada um dos vocábulos, criando tabelas determinísticas de mapeamento (e.g.: todos os nomes “Armando” são sempre mascarados como “Belarmino”).

O facto de serem utilizadas tabelas de mapeamento baseadas em vocábulos fragiliza a privacidade dos dados, já que bastará que um atacante deduza um único registo (com base em outros atributos da base de dados) para que, iterativa e sucessivamente, a tabela de mapeamento fique comprometida.

Id	Nome original	Nome mascarado
1	JOAO MARIA SANTOS DA SILVA FERREIRA	VALTER ISABEL BRITO RIBEIRO MAIA PINHEIRO
2	ANA MARIA FERREIRA JESUS CAMPOS	SORAIA ISABEL PINHEIRO SOARES ALVES
3	ARMANDO MASCARENHAS	BELARMINO SILVA
4	ASDRUBAL JOSE TELES CAMPOS SILVA	CARLOS MANUEL FERREIRA ALVES MAIA
5	MARIA RITA MENDES ALMEIDA	ISABEL SOFIA PEREIRA WELLINGTON

Tabela 8 - *Shuffling*: exemplo de tabela de mapeamento baseado em vocábulos

Analisando a Tabela 8, é possível constatar que, caso o registo n.º 1 seja comprometido, ficarão também parcialmente comprometidos os registos n.º 2, 4 e 5.

O documento “*Oracle – Data Masking Best Practices*” [40] explora em detalhe a aplicação de técnicas determinísticas em bases de dados múltiplas, defendendo o seguinte postulado (Fig. 4):

F = função determinística de *data masking* que preconiza

$$y = F\ (x),$$

não sendo possível aplicar a função inversa, i.e.:

$$x = F^{-1}(y)$$

Fig. 4 - Formulação de uma função determinística de *data masking*

No entanto, a utilização de funções irreversíveis (e.g. substituição de uma *string* pelo respetivo *hash* SHA256) não deverá ser considerada suficiente para garantir a segurança dos dados, já que um atacante poderá aplicar de forma repetida (via *brute--force*) essa mesma função a diferentes palavras/frases de um dicionário, potencialmente deduzindo a tabela de mapeamento.

Uma forma possível de mitigar este ataque passará pela implementação nas funções determinísticas de um fator de entropia suficientemente complexo (e.g. *salt alfanumérico de 20 ou mais caracteres)*.

Outra alternativa de segurança será a adoção de tabelas de mapeamento aleatório, construídas dinamicamente à medida que vão sendo mascarados os registos:

Id	SubId	Nome original	Nome mascarado (após aplicação de função para obtenção de nome aleatório)
1	1.1	JOAO	CARLOS
	1.2	MARIA	ISA
	1.3	SANTOS	CORREIA
	1.4	DA	MATOS
	1.5	SILVA	PEREIRA
	1.6	FERREIRA	ANTUNES
2	2.1	ANA	MANUELA
	2.2	MARIA	JULIA
	2.3	FERREIRA	ARISTIDES
	2.4	JESUS	MENDES
	2.5	CAMPOS	SOUSA
3	3.1	ARMANDO	RICARDO
	3.2	MASCARENHAS	PONTES
4	4.1	ASDRUBAL	RICARDO
	4.2	JOSE	JORGE
	4.3	TELES	MATIAS
	4.4	CAMPOS	PEREIRA
	4.5	SILVA	SANTOS
5	5.1	MARIA	CARLA
	5.2	RITA	MARIA
	5.3	MENDES	VIEIRA
	5.4	ALMEIDA	SILVA

Tabela 9 - *Shuffling*: exemplo de tabela de mapeamento aleatório

Ao contrário da Tabela 8 (onde o nome Armando era sempre codificado como Belarmino), os nomes na Tabela 9 são gerados aleatoriamente, registo a registo.

Quando concluída, esta tabela de mapeamento aleatório deverá ser utilizada no processo de *data masking* de todas as aplicações que partilhem a mesma informação.

Estas tabelas deverão ser destruídas após conclusão de todo o processo de *data masking*, evitando-se que o processo seja passível de reversão.

4.6 *Shuffling* - Ataques por tabela de frequências

Não obstante, o ataque anteriormente referido poderá ser bem-sucedido mesmo sem que seja necessário deduzir previamente qualquer registo, bastando para tal que o atacante utilize uma tabela de frequências: à semelhança das técnicas de análise e descodificação de mensagens adotadas em métodos criptográficos não digitais (Fig. 5Fig. 6 e Fig. 6), também neste caso poderá será possível calcular a frequência dos dados originais, e, a partir daí, proceder à descodificação da tabela de mapeamento.

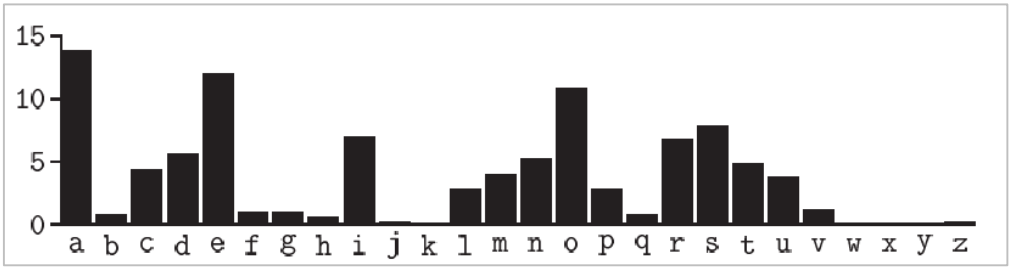


Fig. 5 - Histograma dos caracteres portugueses [41]

a	b	c	d	e	f	g	h	i	j	k	l	m
13.8	0.9	4.5	5.6	12.0	1.0	1.2	0.6	7.0	0.3	0.0	2.8	4.1
n	o	p	q	r	s	t	u	v	w	x	y	z
5.3	10.8	2.9	0.8	6.9	7.8	4.9	3.8	1.3	0.0	0.2	0.0	0.3

Fig. 6 - Percentagens de frequências dos caracteres portugueses [41]

A título de exemplo foram analisadas as ocorrências de primeiros e últimos nomes de uma tabela de clientes *[vd. 5.2 Cenário 2 - Base de Dados de Clientes]*, tendo-se obtido as seguintes percentagens de frequência (Fig. 7 e Fig. 8).

De realçar a ocorrência do nome “LDA” nos “últimos nomes” mais frequentes, a qual corresponde a uma terminação usual do nome comercial de muitas empresas.

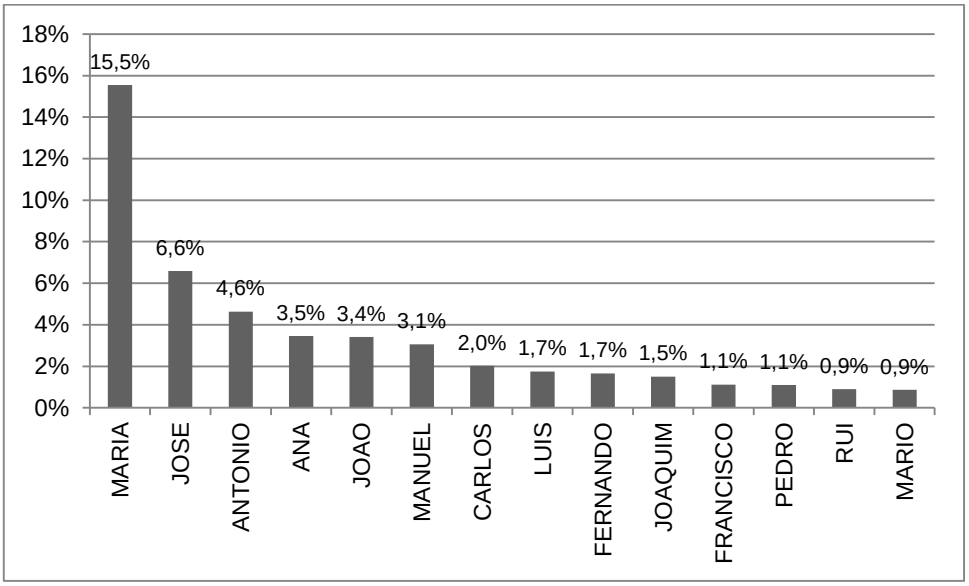


Fig. 7 - Histograma do "primeiro nome" de uma tabela de clientes

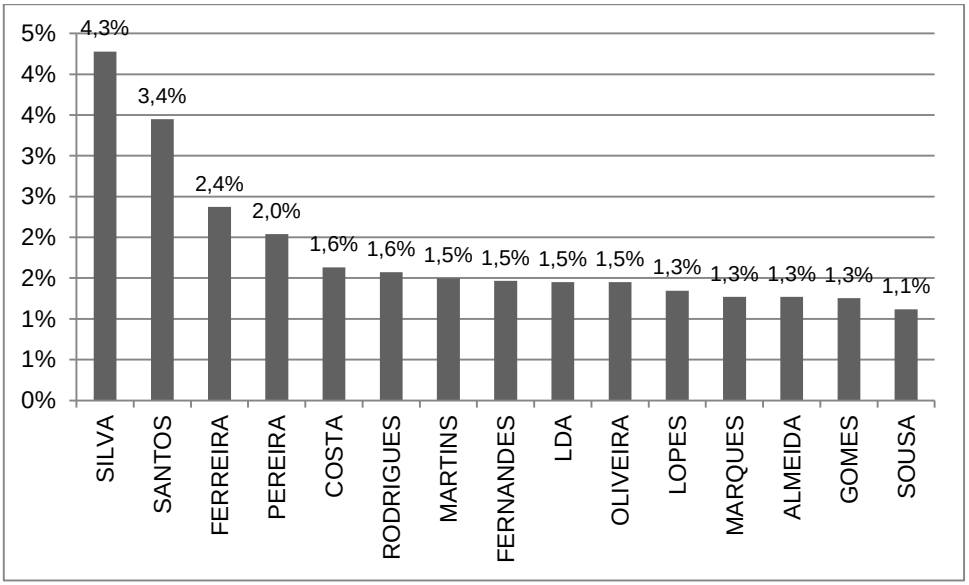


Fig. 8 - Histograma do "último nome" de uma tabela de clientes

Por último, se forem considerados apenas os registos de clientes cujo NIF inicia pelo algarismo 5 (a gama de NIF 5xx é exclusiva de Pessoas Coletivas [42]), é possível constatar que a terminação "LDA" ocorre em 73,1% dos casos.

5. Casos de estudo

De forma a potenciar o conhecimento adquirido nas temáticas da anonimização, de-identificação e pseudonimização, e respetivo alinhamento com o GDPR, foram contactadas diferentes empresas nacionais solicitando uma colaboração prática neste trabalho.

As respostas positivas permitiram desenvolver e explorar quatro cenários com objetivos funcionais distintos, combinando a necessidade de mitigação dos riscos de quebra de privacidade com a preservação da qualidade e veracidade dos dados.

Além da disponibilização de informação para cada caso de estudo, foi solicitada uma reunião prévia para levantamento de requisitos de anonimização, assim como a disponibilidade para uma discussão final dos resultados alcançados.

Por motivos de confidencialidade não serão divulgados nomes ou outras informações relevantes sobre as referidas empresas.

5.1 Cenário 1 - Base de Dados de Candidatos (Processo de Recrutamento de Recursos Humanos)

Os dados analisados neste cenário foram fornecidos por indivíduos externos a uma determinada Empresa, no âmbito da apresentação de candidaturas espontâneas de emprego através de um formulário *online*.

Depois de processados pela aplicação informática, os dados foram armazenados na base de dados do Departamento de Recursos Humanos da Empresa.

Estes dados são utilizados pelo referido Departamento para identificação de candidatos em processos de recrutamento e seleção, estando o acesso a esta base de dados sujeito a mecanismos de autenticação e autorização intrínsecos a uma determinada aplicação informática da Empresa.

5.1.1 Objetivos

A área de comunicação da Empresa demonstrou interesse em aceder regularmente aos referidos dados de candidatos, visando aferir “indicadores públicos de interesse”, desenvolver estudos internos de mercado e posicionamento da marca, assim como realizar análises cruzadas com informações disponibilizadas por outras fontes externas à empresa.

Para identificação de requisitos foi realizada uma primeira reunião com a área de comunicação, a partir da qual foi possível confirmar que os dados poderão ser parcialmente de-identificados, sem que tal implique uma diminuição do valor dos estudos; foi apenas deixada a ressalva para que, dentro do possível, os dados anonimizados continuem a representar o mais fielmente possível a realidade original.

Por outro lado, uma vez que deixará de ter qualquer controlo sobre os dados, o Departamento de Recursos Humanos considerou insuficiente a realização de um processo de de-identificação, pelo que (dentro do possível) os dados deverão ser efetivamente anonimizados, tornando o processo irreversível.

5.1.2 Metodologia Proposta

Numa primeira análise à estrutura da base de dados foi possível constatar que alguns dos atributos teriam de ser necessariamente omitidos, já que possuíam um forte potencial de identificação unívoca de cada candidato. Por outro lado, o atributo “média final de curso” evidenciou-se como especialmente relevante e sensível.

Após sucessivas iterações e discutidas as categorias de dados disponíveis, foi proposta a minimização do risco de re-identificação através da aplicação de um conjunto de técnicas de anonimização por generalização, nomeadamente *K-Anonymity* e *L-Diversity* [43].

Face aos objetivos propostos foram definidas as seguintes variáveis de partida:

- limitar a 20% a proporção de registos válidos descartados (potenciais *outliers*);
- maximizar a qualidade dos atributos espaço temporais: “distrito” e “data de candidatura” (minimizando a generalização);
- destacar o atributo “mediaCurso” como “sensível” (“*sensitive*”), de forma a que o mesmo não seja generalizado ou suprimido, assegurando a heterogeneidade deste atributo dentro do mesmo grupo equivalente.

5.1.3 Análise e tratamento prévio dos dados

Os dados foram recebidos em formato CSV após extração direta da tabela da base de dados original, tendo uma primeira análise revelado as seguintes informações (Tabela 10):

Nº de registos	3561
Nº de atributos	19
Lista de atributos	ID, Nome, BI, DataNascimento, Email, Telefone, DataRecepcao, Título, Sexo, Distrito, SituaçãoAtual, AnoInícioProfissão, AnosFormação, Escola de Formação, Curso de Formação, AnoInicioCurso, MédiaCurso, CAP e PDF_CurriculumVitae

Tabela 10 - Cenário 1 – Caraterização da base de dados de origem

Uma vez que o processo de exportação para CSV foi realizado diretamente a partir da aplicação dos RH, sem quaisquer parametrizações, foi necessário proceder previamente a uma limpeza técnica dos dados, de modo a garantir o seu correto tratamento:

- correção de erros de *charset*;
- conversão de datas para formato aaaa-mm-dd [ISO 8601];
- conversão do separador de coluna (vírgula => ponto e vírgula);
- conversão de separador de decimais (ponto => vírgula);

Uma parte significativa dos campos havia sido tipificada no formulário *web* como texto livre, sem obedecer a formatos pré-definidos, prejudicando substancialmente o processo de análise e categorização. Exemplos:

- a recolha do campo “Telefone” não obedeceu a qualquer tipificação, permitindo respostas tão heterogéneas como "*351-22-xxxxxxx*", "*35122xxxxxxx*", "*22xxxxxxx*" ou mesmo “*22xxxx (após as 16h)*”
- o campo “MédiaCurso” possuía respostas abertas do tipo “*a frequentar*” ou “*previsão 14*”; foram eliminados todos os registos com formatos não conformes e com valores inferiores a 10 ou superiores a 20.

Constatou-se também que a aplicação não possuía nenhum mecanismo *anti-bots* e permitiu, ao longo de vários anos, a introdução de dados de candidatos sem qualquer validação de sintaxe ou contexto (datas, emails, etc.), inevitavelmente permitindo um significativo conjunto de registos inválidos.

Após limpeza, eliminação de duplicados e de registos inválidos, a base de dados ficou reduzida a um total de 2362 registos.

Por fim foram suprimidos os seis atributos identificadores (ID, Nome, BI, Email, Telefone e PDF_CV) e importada a nova tabela para o *software* de anonimização ARX (Fig. 9 e Fig. 10).

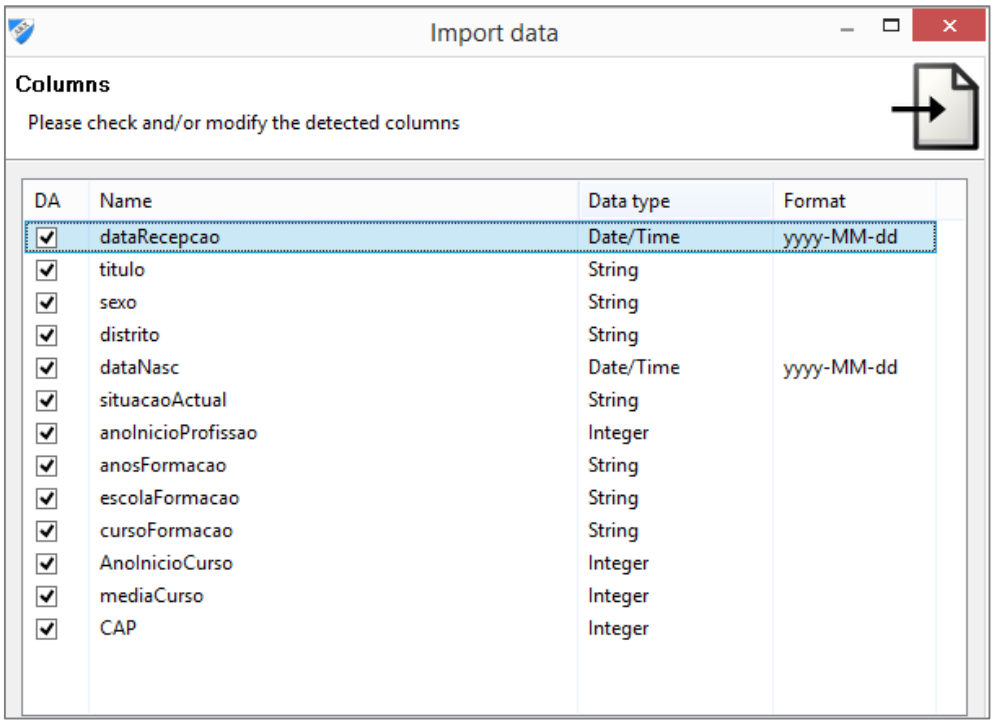


Fig. 9 - Cenário 1 – Importação de dados para o ARX

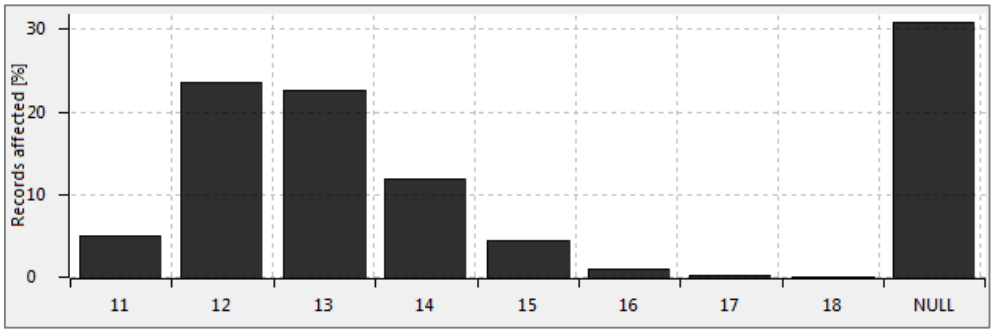


Fig. 10 - Cenário 1 - Histograma do atributo “média de curso” (dados base)

Aferindo os riscos de quebra de privacidade inicial (adotando o modelo previamente apresentado em [*3.2 Riscos*]) obtiveram-se valores próximos de 100% (Fig. 11), situação comprovada pelos seguintes factos:

- o atributo “data de nascimento” permitia identificar univocamente 85% dos registos;

- o atributo “data de receção da candidatura” permitia identificar univocamente 43% dos registos;
- a simples combinação destes dois atributos permitia identificar univocamente 99.9% dos registos.

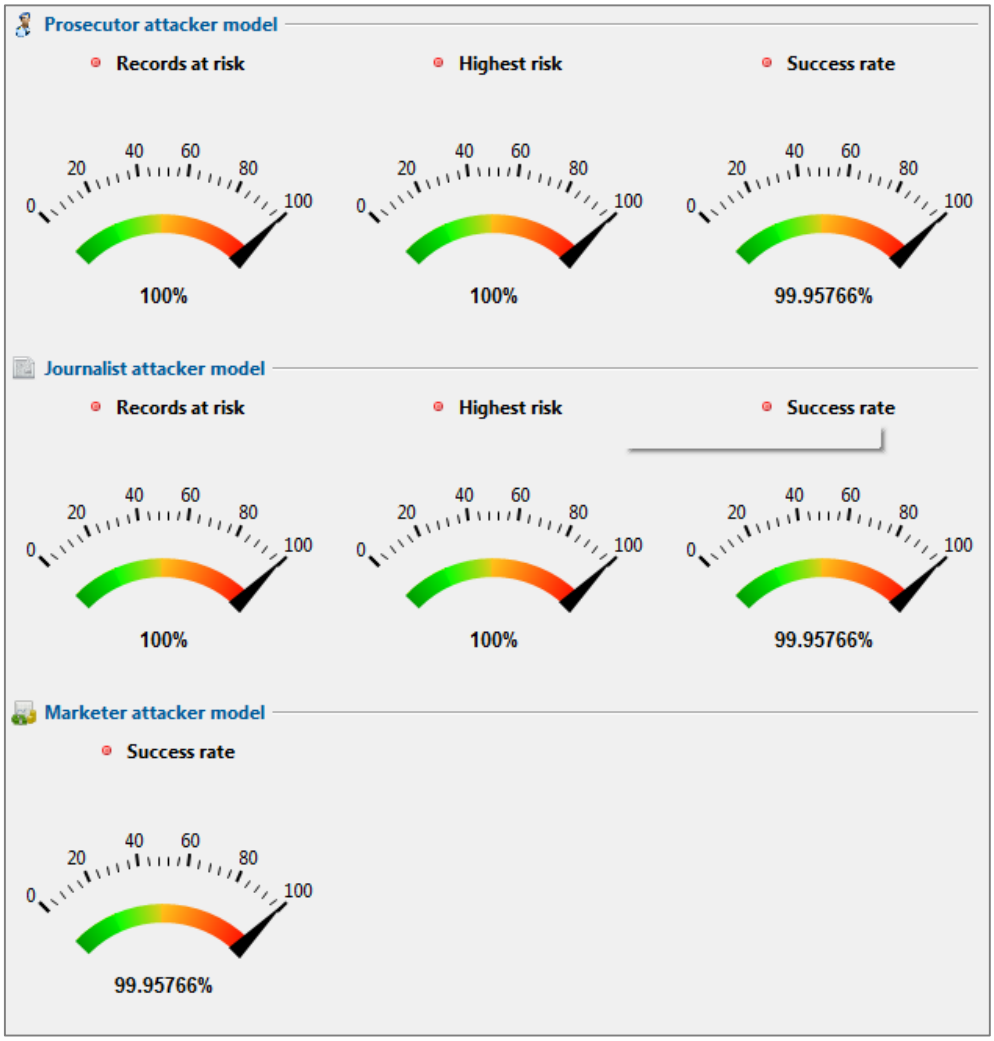


Fig. 11 - Cenário 1 – Riscos de quebra de privacidade (dados base)

De forma a mitigar este risco foi então desenhado um modelo de anonimização no *software* ARX, tendo a proposta inicial de tipificação de dados obedecido ao exposto na Tabela 11:

Atributo	% Reg.	Tipo	Técnica de Generalização	Níveis
dataRecepcao	100%	*Quasi-identifier*	Intervalos de datas	0-5
titulo	66%	*Quasi-identifier*	Máscara de caracteres	0-5
sexo	99%	*Quasi-identifier*	Máscara de caracteres	0-1
distrito	95%	*Quasi-identifier*	Agrupar em classes Máscara de caracteres	0-20
dataNasc	99%	*Quasi-identifier*	Intervalos de datas	0-7
situacaoActual	100%	*Quasi-identifier*	Máscara de caracteres	0-12

Atributo	% Reg.	Tipo	Técnica de Generalização	Níveis
anoInicioProfissao	100%	*Quasi-identifier*	Intervalos de datas	0-5
anosFormacao	99%	*Quasi-identifier*	Agrupar em classes Máscara de caracteres	0-4
escolaFormacao	74%	*Quasi-identifier*	Máscara de caracteres	0-12
cursoFormacao	69%	*Quasi-identifier*	Máscara de caracteres	0-11
anoInicioCurso	96%	*Quasi-identifier*	Intervalos de datas	0-4
mediaCurso	69%	***Sensitive***	N / A	N / A
CAP	52%	*Quasi-identifier*	Máscara de caracteres	0-4

Tabela 11 - Cenário 1 – Tipificação de dados e técnicas de generalização propostas

A hierarquia de generalização dos atributos “distrito” e “anosFormacao” (escolaridade) foi testada com as opções “máscara” e “agrupamento em classes” (e.g. distritos agrupados pelas regiões centro, norte e sul), sem que tal tenha provocado alterações significativas nos resultados.

A título de exemplo é apresentado na Fig. 12 um excerto da hierarquia de generalização adotada para o atributo “dataRecepcao”, e na Tabela 12 as parametrizações iniciais do modelo de privacidade.

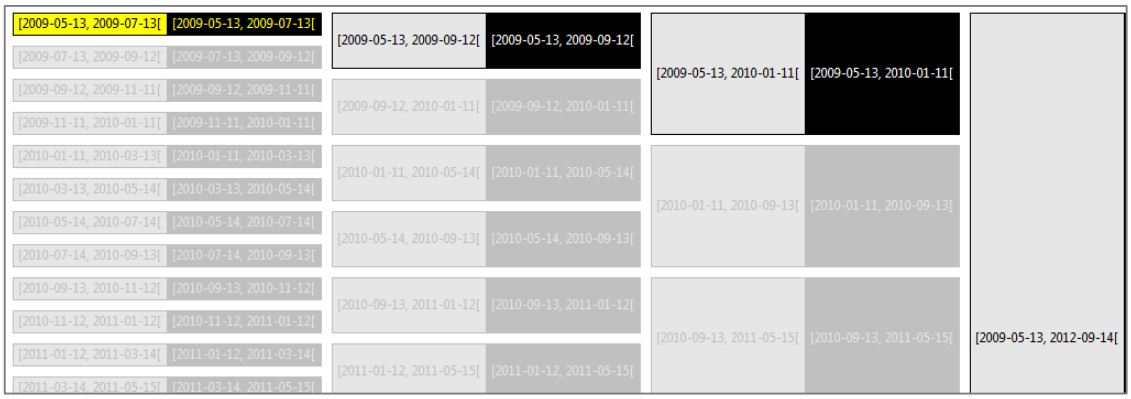


Fig. 12 - Cenário 1 – Hierarquia de Generalização do atributo “dataRecepcao” – excerto

Parâmetro	Valor
Supression Limit	20%
Attribute Weights	*“dataRecepcao”* = 1.0 “distrito” = 1.0
Privacy Model	*K-Anonymity* com K=3 *L-Diversity* com L=2

Tabela 12 - Cenário 1 – Parâmetros do modelo de privacidade

5.1.4 Resultados

Os primeiros resultados obtidos não foram considerados satisfatórios, face à generalização total de muitos dos atributos, assim como devido à significativa perda de qualidade dos dados indicados como prioritários.

Cenário 1 – modelo 1 13 atributos; *K-Anonymity*=3; *L-Diversity*=2	
% de registos suprimidos	12.5%
Nº de classes Equivalentes	138
Transformação aplicada	[5,0,0,0,7,4,5,4,12,11,4,4] [2]
Generalização do atributo "dataRecepcao"	Nível 5 (generalização máxima), implicando a total supressão deste dado
Generalização do atributo "distrito"	Nível 0 (generalização mínima), permitindo manter toda a informação deste dado
Risco médio de re-identificação	6.63%

Tabela 13 - Cenário 1 – Resultados gerais após anonimização – modelo 1

Assim, foi promovida uma nova reunião com a área de Comunicação da Empresa, na qual foi possível discutir e negociar quais os dados efetivamente considerados críticos e relevantes para as análises de mercado. Foi reafirmada a extrema importância dos atributos espaço temporais ("distrito" e "data de candidatura"), colocando-se em segundo nível os atributos "idade" e "sexo"; face aos riscos de quebra de privacidade, os restantes atributos foram admitidos como descartáveis.

Com base nesta informação foi desenvolvido um novo modelo, utilizando agora apenas cinco atributos (Tabela 14):

Atributo	% Reg.	Tipo	Técnica de Generalização	Níveis
dataRecepcao	100%	*Quasi-identifier*	Intervalos de datas	0-5
sexo	99%	*Quasi-identifier*	Máscara de caracteres	0-1
distrito	95%	*Quasi-identifier*	Agrupar em classes Máscara de caracteres	0-20
dataNasc	99%	*Quasi-identifier*	Intervalos de datas	0-5
mediaCurso	69%	***Sensitive***	Intervalos de valores	N / A

Tabela 14 - Cenário 1 – Tipificação de dados e técnicas de generalização propostas – modelo 2

[2] A "transformação aplicada" explicita o nível de generalização aplicado a cada atributo, obedecendo à mesma ordem e escala apresentadas na Tabela 11. O nível 0 indica que o atributo foi mantido inalterado (sem generalização), enquanto o nível máximo significa a perda da informação do atributo.

Das diversas combinações possíveis, foi selecionada a transformação que minimizava a generalização dos atributos "dataRecepcao" e "distrito", ao mesmo tempo que eram aplicados os restantes pressupostos de anonimização. Os resultados estão patentes na Tabela 15 e nas Fig. 13 e Fig. 14:

Cenário 1 – modelo 2 5 atributos; *K-Anonymity*=3; *L-Diversity*=2	
% de registos suprimidos	18.2%
Nº de classes Equivalentes	432
Transformação aplicada	[1,0,0,4] [3]
Generalização do atributo "dataRecepcao"	Nível 1: a data é apresentada como um intervalo de 2meses
Generalização do atributo "distrito"	Nível 0 (generalização mínima), permitindo manter toda a informação deste dado
Risco médio de re-identificação	11.45%

Tabela 15 - Cenário 1 – Resultados gerais após anonimização – modelo 2

Adicionalmente foi garantida a generalização mínima do atributo "sexo", ao contrário do atributo "dataNascimento" (que foi generalizado na sua quase totalidade).

	dataRecepcao	sexo	distrito	dataNasc	mediaCurso
1893	[2011-01-12, 2011-03-14[	M	Viana do Castelo	[1948-06-08, 1995-09-19[	14
1894	[2011-05-15, 2011-07-15[	M	Viana do Castelo	[1948-06-08, 1995-09-19[	11
1895	[2011-05-15, 2011-07-15[	M	Viana do Castelo	[1948-06-08, 1995-09-19[	13
1896	[2011-05-15, 2011-07-15[	M	Viana do Castelo	[1948-06-08, 1995-09-19[	15
1897	[2012-01-13, 2012-03-14[	M	Viana do Castelo	[1948-06-08, 1995-09-19[	12
1898	[2012-01-13, 2012-03-14[	M	Viana do Castelo	[1948-06-08, 1995-09-19[	NULL
1899	[2012-01-13, 2012-03-14[	M	Viana do Castelo	[1948-06-08, 1995-09-19[	NULL
1900	[2010-03-13, 2010-05-14[	F	Vila Real	[1948-06-08, 1995-09-19[	12
1901	[2010-03-13, 2010-05-14[	F	Vila Real	[1948-06-08, 1995-09-19[	13
1902	[2010-03-13, 2010-05-14[	F	Vila Real	[1948-06-08, 1995-09-19[	13
1903	[2009-07-13, 2009-09-12[	M	Viseu	[1948-06-08, 1995-09-19[	12
1904	[2009-07-13, 2009-09-12[	M	Viseu	[1948-06-08, 1995-09-19[	13
1905	[2009-07-13, 2009-09-12[	M	Viseu	[1948-06-08, 1995-09-19[	13
1906	[2009-07-13, 2009-09-12[	M	Viseu	[1948-06-08, 1995-09-19[	NULL
1907	[2009-07-13, 2009-09-12[	M	Viseu	[1948-06-08, 1995-09-19[	NULL
1908	[2009-07-13, 2009-09-12[	M	Viseu	[1948-06-08, 1995-09-19[	NULL
1909	[2009-09-12, 2009-11-11[	M	Viseu	[1948-06-08, 1995-09-19[	11
1910	[2009-09-12, 2009-11-11[	M	Viseu	[1948-06-08, 1995-09-19[	11
1911	[2009-09-12, 2009-11-11[	M	Viseu	[1948-06-08, 1995-09-19[	12
1912	[2009-09-12, 2009-11-11[	M	Viseu	[1948-06-08, 1995-09-19[	12
1913	[2009-09-12, 2009-11-11[	M	Viseu	[1948-06-08, 1995-09-19[	14
1914	[2009-09-12, 2009-11-11[	M	Viseu	[1948-06-08, 1995-09-19[	NULL
1915	[2010-03-13, 2010-05-14[	M	Viseu	[1948-06-08, 1995-09-19[	12

Fig. 13 - Cenário 1 – Excerto da tabela de dados anonimizados – modelo 2

[3] A "transformação aplicada" explicita o nível de generalização aplicado a cada atributo, obedecendo à mesma ordem e escala apresentadas na Tabela 14. O nível 0 indica que o atributo foi mantido inalterado (sem generalização).

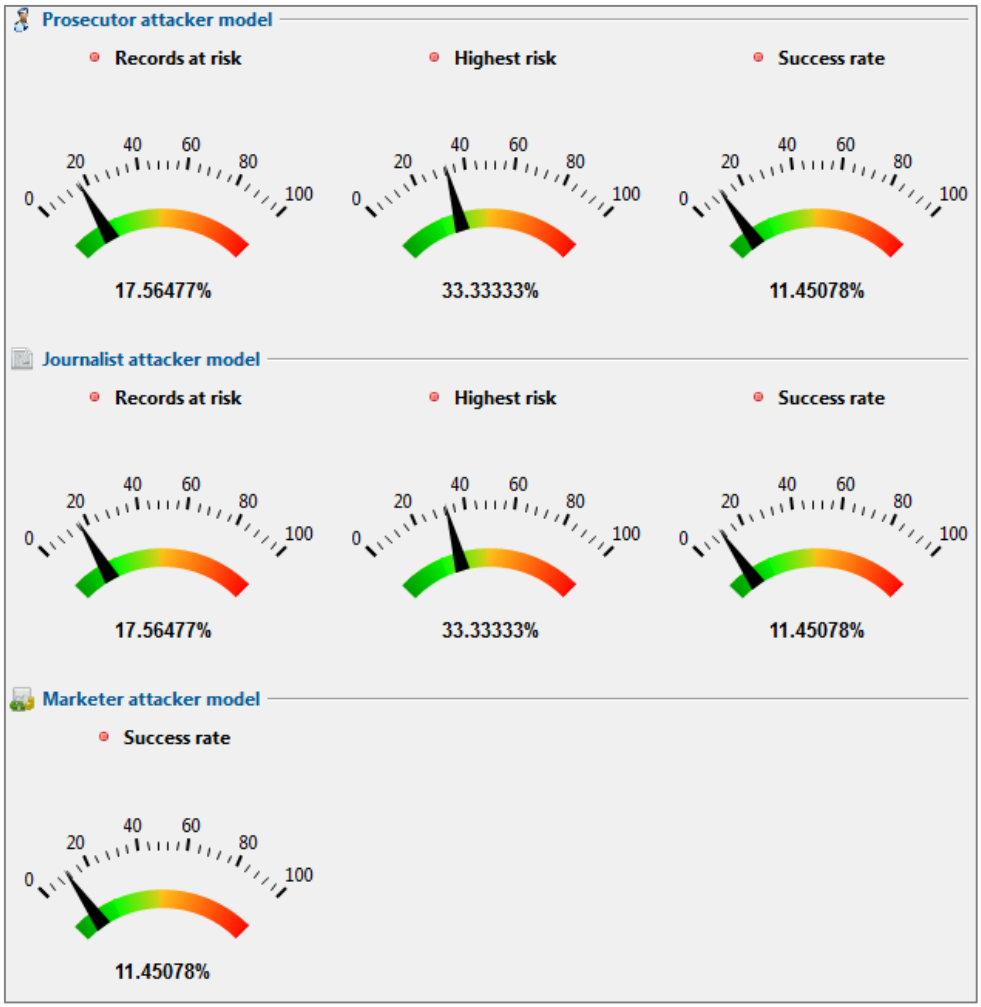


Fig. 14 - Cenário 1 – Riscos de quebra de privacidade – modelo 2

Os resultados alcançados com este segundo modelo foram discutidos e validados com a Empresa, tendo sido apenas lamentada a perda de informação relativa à idade dos candidatos. Verificou-se que mesmo reduzindo o parâmetro *K-Anonymity* para K=2 (aumentando o risco de re-identificação), o nível de generalização desse atributo não sofreria qualquer alteração. Por outro lado constatou-se que a idade da maioria dos candidatos estava inserida num curto intervalo de oito anos, pelo que a manutenção deste atributo revelou-se de menor importância, face às vantagens de ser maximizada a privacidade da informação.

Foi também discutida a diversidade dos dados, uma vez que foram suprimidos 18% dos registos (ainda assim um valor abaixo do limite inicialmente estabelecido de 20%) por não se enquadrarem em nenhum dos grupos equivalentes estabelecidos pelo modelo. Assumiu-se que essa é uma consequência do processo de anonimização, mas que tal não deverá implicar diferenças significativas nos estudos de mercado que se pretendem realizar.

5.2 Cenário 2 - Base de Dados de Clientes

A Empresa possui uma base de dados de clientes onde constam informações contratuais, de faturação e consumo dos diferentes clientes, na qual estão armazenados atualmente milhares de registos. Esta base de dados é utilizada por diversas aplicações informáticas, integradas entre si, sendo composta por diversas tabelas num modelo relacional.

5.2.1 Objetivos

Pretende-se assegurar que os ambientes de testes (DEV e QLD) possuam dados reais (de produção) de-identificados, de forma a dar seguimento à implementação de uma estratégia de "*Privacy by Design*" [1].

Dentro do tecnicamente possível estes dados deverão representar a realidade do ambiente produtivo e, obrigatoriamente, respeitar toda a sintaxe e tipologia dos dados de origem. Poderão ser suprimidos até 50% dos registos de origem, embora sendo desejável manter o máximo de diversidade.

5.2.2 Metodologia Proposta

Embora tratando-se de um processo de de-identificação dos dados de produção, para utilização em ambientes de testes, considerou-se fundamental utilizar uma abordagem baseada em técnicas de Generalização e *Randomization*, de forma a ser possível estimar os riscos de quebra de privacidade, o que não seria possível caso fossem adotadas técnicas de pseudonimização ou *data masking* [44] *[vd.3.2 Pseudonimização].*

De forma a respeitar os objetivos indicados foram desenvolvidos dois modelos de privacidade:

- ➔ De-identificação por "*Differential Privacy*", com o objetivo de minimizar os riscos de quebra de privacidade;

- ➔ De-identificação por "*K-Anonymity*", com o objetivo de preservar o máximo de diversidade dos dados.

Por se tratar de uma base de dados composta por diferentes tabelas, o processo de de-identificação implicará:

1. Construção de uma tabela única [tuplos];
2. Aplicação de técnicas de anonimização;
3. Recálculo de valores individuais, com base nos intervalos criados pela generalização (de modo a respeitar o *datatype* da base de dados);
4. Reconstrução das tabelas de origem, com base nos dados de-identificados;
5. Por fim, todos os campos não abrangidos pelo processo deverão ser mascarados (e.g. informações relativas ao cliente) ou gerados a partir dos dados já processados (e.g. datas de consumos e faturação).

5.2.3 Análise e tratamento prévio dos dados

Não sendo fundamental garantir a presença da totalidade dos registos na base de dados de-identificada, o tratamento de dados focou-se na deteção e eliminação de registos "não conformes", evitando assim a morosa tarefa de correção individualizada de registos erróneos.

Verificou-se que a informação relevante para o processo de de-identificação encontrava--se distribuída por quatro tabelas relacionadas entre si, com a estrutura indicada na Tabela 16.

Base de dados de clientes – Tabela "CLIENTES"	
Nº de registos	14679
Nº de atributos	7
Lista de atributos	CodigoCliente, Nome, Morada, Telefone, NIF, Email, TipoCliente
Base de dados de clientes – Tabela "CONTRATOS"	
Nº de registos	14854
Nº de atributos	4
Lista de atributos	CodigoCliente, ID, DataContrato, Tarifario
Base de dados de clientes – Tabela "REGISTO_CONSUMOS"	
Nº de registos	93854
Nº de atributos	4
Lista de atributos	ID, DataRegisto, ValorConsumido, DifConsumAnterior
Base de dados de clientes – Tabela "FATURAÇÃO"	
Nº de registos	140540
Nº de atributos	3
Lista de atributos	ID, ValorFatura, PeriodoFaturado

Tabela 16 - Cenário 2 – Caraterização da base de dados de origem

De forma a permitir a aplicação das técnicas de anonimização à base de dados original os registos foram convertidos para uma tabela única de tuplos, utilizando os atributos “CodigoCliente” e “ID” como atributos de conexão.

Face aos objetivos propostos admitiu-se que cada cliente deveria possuir um único tuplo, suprimindo-se todos os dados pessoais.

Não foi considerada relevante a preservação do historial de consumos e faturação ao longo do tempo. Para tal foi necessário realizar uma operação de conversão de todos os registos cronológicos de faturação e consumos, de forma a obter valores médios diários por cliente. Estes dados foram calculados e preenchidos em pelo menos 70% dos clientes, ficando marcados como omissos nos restantes. De notar que, face à lógica aplicacional existente, estes valores podem ser inferiores a zero.

A tabela de tuplos com 14679 registos foi posteriormente importada para o *software* ARX (Fig. 15 e Fig. 16), considerando-se um total de seis atributos.

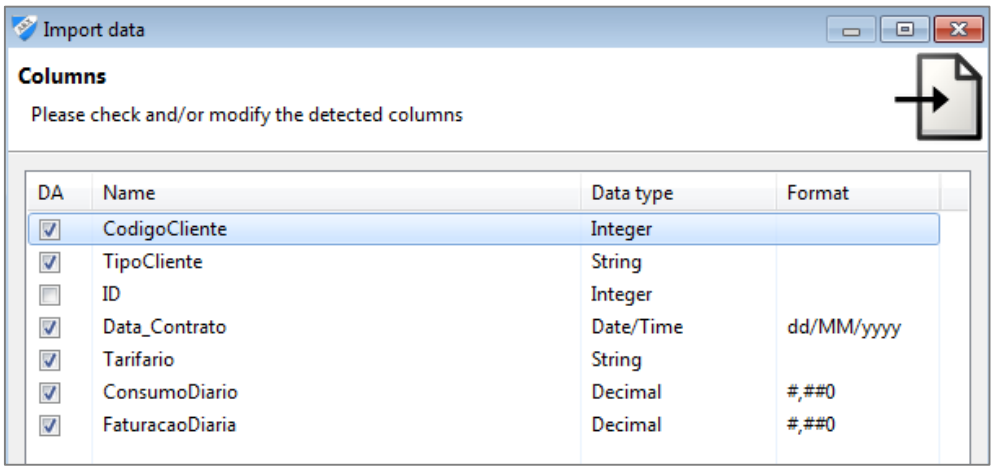


Fig. 15 - Cenário 2 – Importação de dados para o ARX

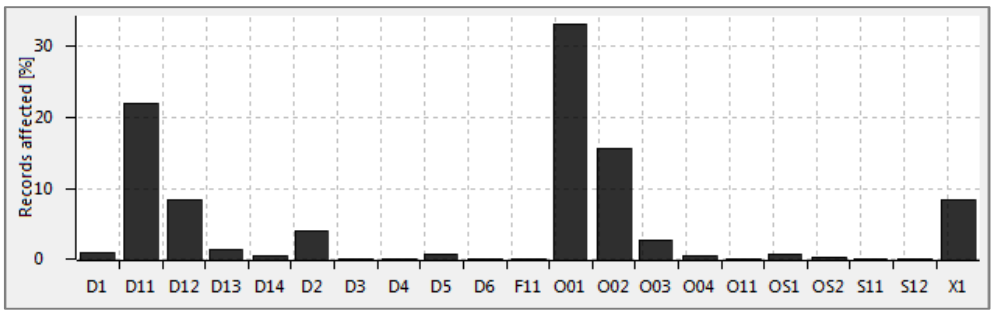


Fig. 16 - Cenário 2 - Histograma do atributo “Tarifario” (dados base)

Mesmo tratando-se de uma base de dados com um volume significativo de registos, a estimativa de risco inicial (sem de-identificação) aponta para valores elevados (Fig. 17).

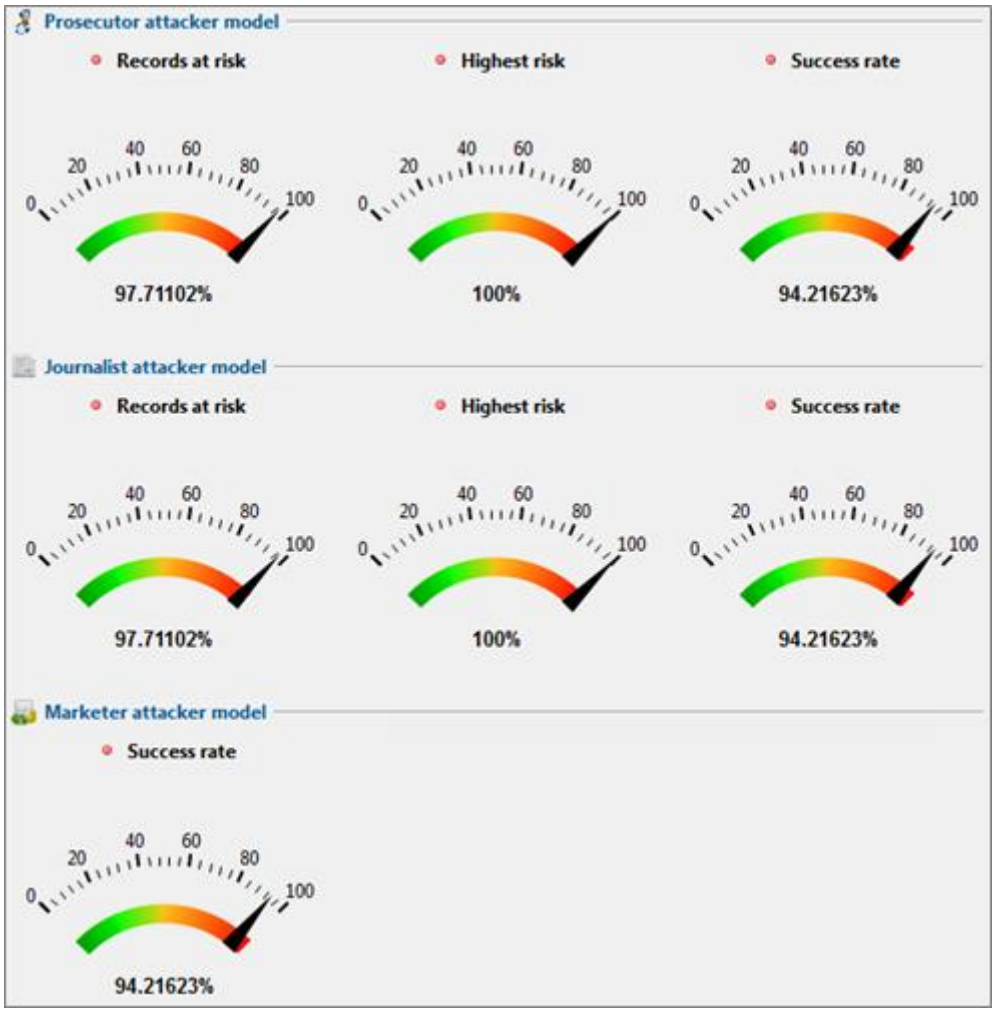


Fig. 17 - Cenário 2 – Riscos de quebra de privacidade (dados base)

São de realçar as seguintes constatações:

- O atributo “Data_Contrato” permitiria identificar univocamente 36% dos clientes;
- A combinação dos atributos “Data_Contrato” e “FaturacaoDiaria” permitiria identificar univocamente 90% dos clientes;

De forma a cumprir com os objetivos inicialmente traçados foi desenhado um modelo de de-identificação no *software* ARX, tendo a proposta inicial de tipificação de dados obedecido ao exposto na Tabela 17:

Atributo	% Reg.	Tipo	Técnica de Generalização	Níveis
CodigoCliente	100%	*Insensitive*	--	--
TipoCliente	100%	*Quasi-identifier*	Agrupar em classes	0-1
Data_Contrato	100%	*Quasi-identifier*	Intervalos de datas	0-4
Tarifario	100%	*Quasi-identifier*	Agrupar em classes	0-1
ConsumoDiario	72%	*Quasi-identifier*	Intervalos de valores	0-4
FaturacaoDiaria	74%	*Quasi-identifier*	Intervalos de valores	0-6

Tabela 17 - Cenário 2 – Tipificação de dados e técnicas de generalização propostas

A título de exemplo é apresentado na Fig. 18 um excerto da hierarquia de generalização adotada para o atributo “ConsumoDiario”.

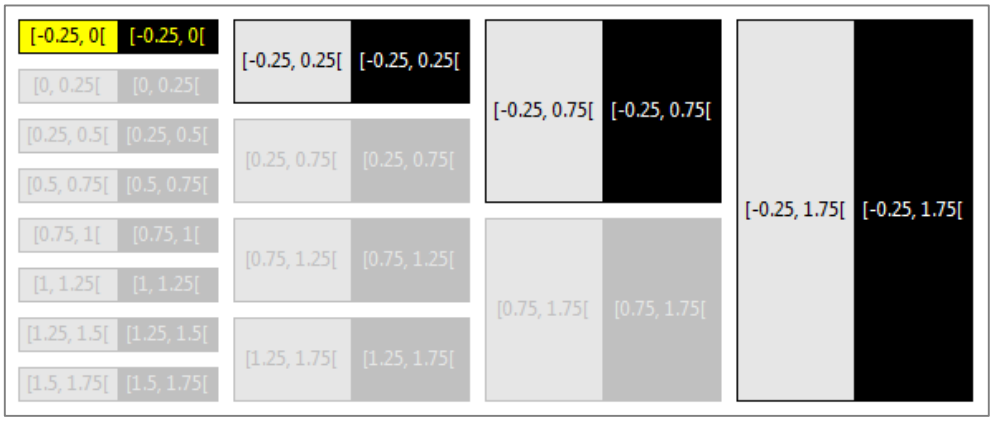


Fig. 18 - Cenário 2 – Hierarquia de Generalização do atributo “ConsumoDiario”

Na Tabela 18 são explicitadas as parametrizações iniciais dos dois modelos de privacidade adotados.

Parâmetro	Modelo 1	Modelo 2
Supression Limit	50%	5%
Attribute Weights	“TipoCliente” = 1.0 “Tarifario” = 1.0 “ConsumoDiario” = 0.2 “FaturacaoDiaria” = 0.2	
Privacy Model	*Differential Privacy* com generalização média	*K-Anonymity* com K=3

Tabela 18 - Cenário 2 – Parâmetros dos modelos de privacidade

5.2.4 Resultados

Os resultados obtidos com o primeiro modelo (*Differential Privacy*) primam pela redução muito significativa dos riscos de quebra de privacidade, embora sacrifiquem a qualidade e utilidade dos dados, conforme apresentado na Tabela 19, Fig. 19 e Fig. 21.

Cenário 2 – modelo 1	
5 *quasi-identifiers*; Diff Privacy [ε=2, Δ=1e-6, *Medium Generalization*]	
% de registos suprimidos	22.7%
Nº de classes Equivalentes	19 (com cerca de 669 registos/classe)
Transformação aplicada	[0,2,0,3,3]
Generalização do atributo "DataContrato"	Nível 2: a data contratual é apresentada como um intervalo de 22 anos
Generalização do atributo "ConsumoDiario"	Nível 3: o consumo diário é apresentado como um intervalo de 100 centésimas
Generalização do atributo "FaturacaoDiaria"	Nível 3: a faturação diária é apresentada como um intervalo de 100 centésimas
Risco médio de re-identificação	0.18%

Tabela 19 - Cenário 2 – Resultados gerais após de-identificação – modelo 1

	OD_CLI	TIP_CLI	Data_Contrato	Tarifario	Consumo Diario	Faturacao Diaria
3973	018	TC091	[07/10/1994, 11/29/2016[	D12	[-0.25, 0.75[	[0, 1[
3974	478	TC091	[07/10/1994, 11/29/2016[	D12	[-0.25, 0.75[	[0, 1[
3975	958	TC091	[07/10/1994, 11/29/2016[	D12	[-0.25, 0.75[	[0, 1[
3976	139	TC091	[07/10/1994, 11/29/2016[	D12	[-0.25, 0.75[	[0, 1[
3977	192	TC091	[07/10/1994, 11/29/2016[	D12	[-0.25, 0.75[	[0, 1[
3978	566	TC091	[07/10/1994, 11/29/2016[	D12	[-0.25, 0.75[	[0, 1[
3979	085	TC091	[07/10/1994, 11/29/2016[	D12	[0.75, 1.75[	[0, 1[
3980	749	TC091	[07/10/1994, 11/29/2016[	D12	[0.75, 1.75[	[0, 1[
3981	751	TC091	[07/10/1994, 11/29/2016[	D12	[0.75, 1.75[	[0, 1[
3982	571	TC091	[07/10/1994, 11/29/2016[	D12	[0.75, 1.75[	[0, 1[
3983	333	TC091	[07/10/1994, 11/29/2016[	D12	[0.75, 1.75[	[0, 1[
3984	227	TC091	[07/10/1994, 11/29/2016[	D12	[0.75, 1.75[	[0, 1[
3985	349	TC091	[07/10/1994, 11/29/2016[	D12	[0.75, 1.75[	[0, 1[

Fig. 19 - Cenário 2 – Excerto da tabela de dados de-identificados - modelo 1

É também perceptível a perda de diversidade dos dados ao analisar o histograma do atributo "Tarifario" (Fig. 20), já que um número significativo de registos "*outliers*" foi suprimido relativamente à versão base (Fig. 16).

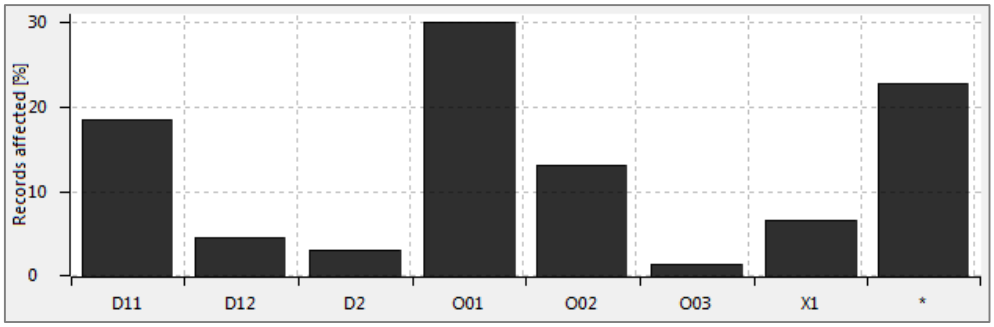


Fig. 20 - Cenário 2 - Histograma do atributo "Tarifario" – modelo 1

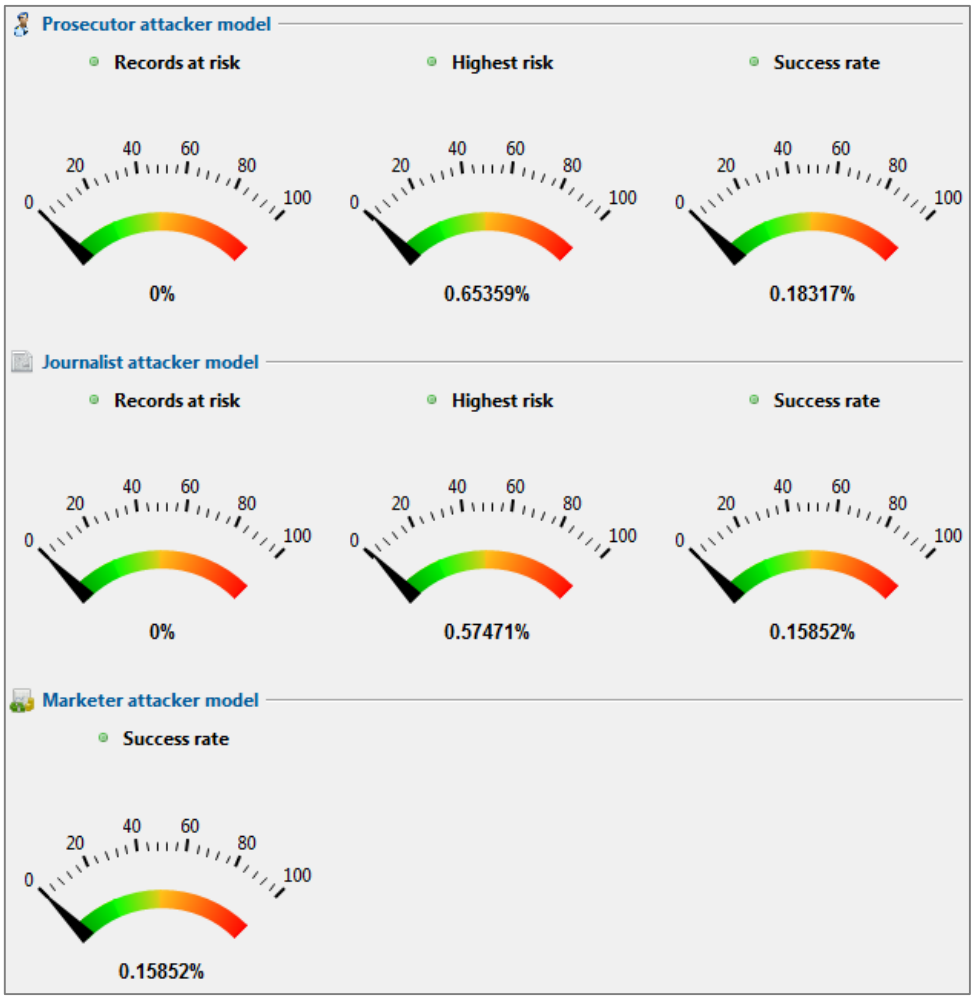


Fig. 21 - Cenário 2 – Riscos de quebra de privacidade – modelo 1

Já no que diz respeito ao segundo modelo, foi possível salvaguardar a privacidade dos dados, mantendo simultaneamente a qualidade e utilidade da informação (Tabela 20).

Cenário 2 – modelo 2 5 *quasi-identifiers*; *K-Anonymity*=3	
% de registos suprimidos	4.7%
Nº de classes Equivalentes	366 (com cerca de 40 registos/classe)
Transformação aplicada	[0,1,0,1,1]
Generalização do atributo "DataContrato"	Nível 1: a data contratual é apresentada como um intervalo de 11 anos
Generalização do atributo "ConsumoDiario"	Nível 1: o consumo diário é apresentado como um intervalo de 25 centésimas
Generalização do atributo "FaturacaoDiaria"	Nível 1: a faturação diária é apresentada como um intervalo de 25 centésimas
Risco médio de re-identificação	2.61%

Tabela 20 - Cenário 2 – Resultados gerais após de-identificação – modelo 2

Uma vez que foi adotado um fator *K-Anonymity*=3 a probabilidade de re-identificação será 1/K, ou seja, 33,33%. Por outro lado, e admitindo que um potencial atacante não possui informação além da existente nestes resultados, a estimativa de risco médio de re-identificação será bastante inferior: 2.6% (Fig. 22).

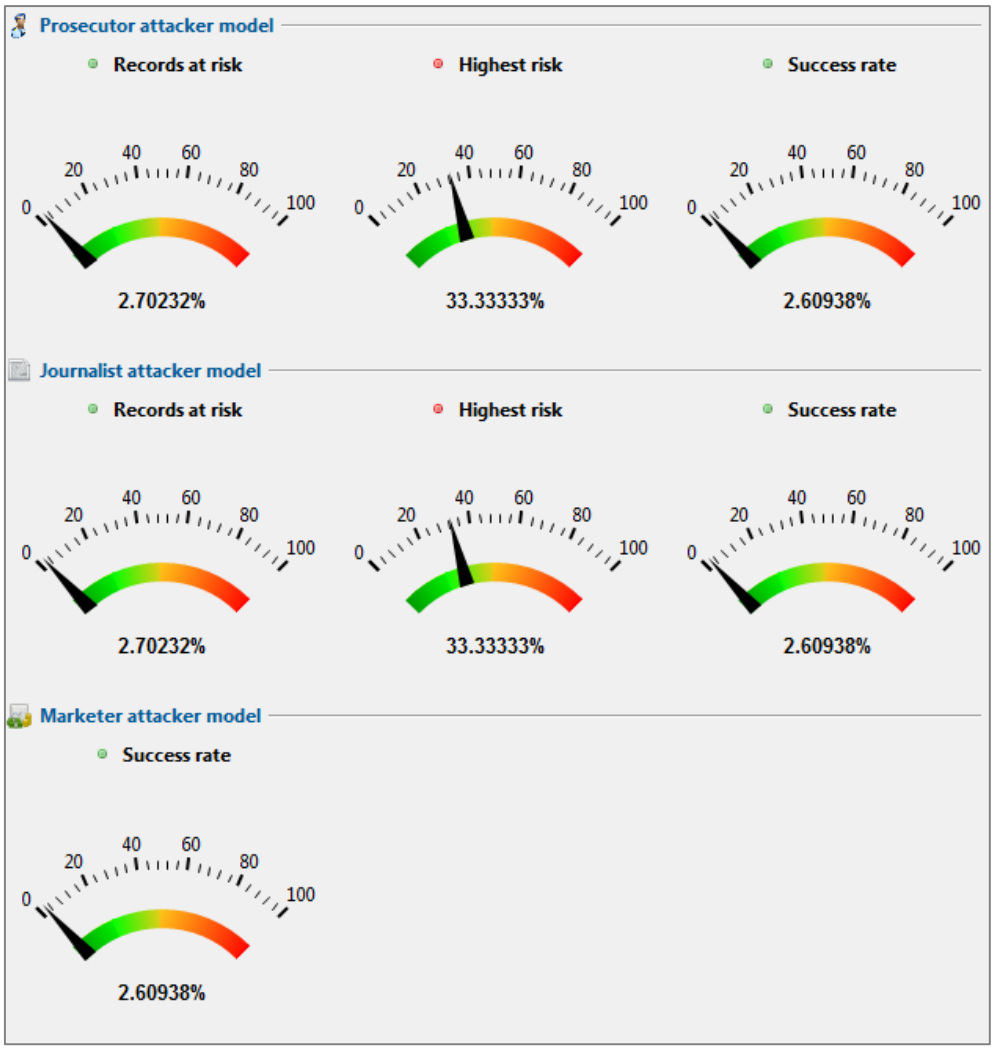


Fig. 22 - Cenário 2 – Riscos de quebra de privacidade – modelo 2

A diversidade dos dados de-identificados (Fig. 24) manteve-se próxima da base de dados original, o que é comprovado pelo histograma da Fig. 23.

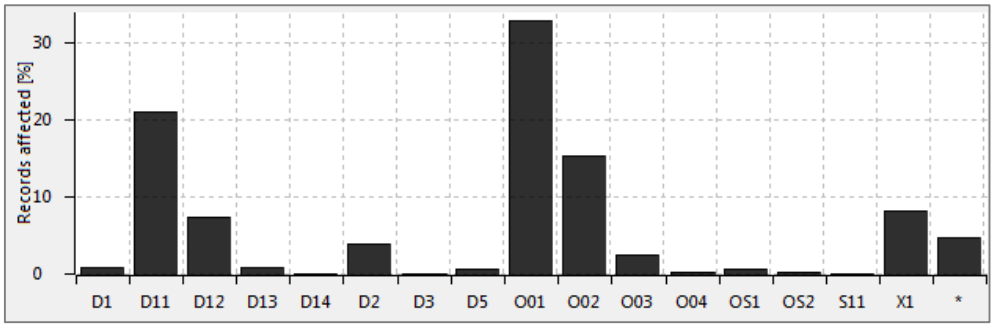


Fig. 23 - Cenário 2 - Histograma do atributo “Tarifario” – modelo 2

	)D_...	TIP_CLI	Data_Contrato	Tarifario	Consumo Diario	Saturacao Diaria
13944	21087	TC091	[09/22/2005, 11/29/2016[	O03	>=1.75	[1, 1.25[
13945	30135	TC091	[09/22/2005, 11/29/2016[	O03	>=1.75	[1, 1.25[
13946	95379	TC091	[09/22/2005, 11/29/2016[	O03	>=1.75	[1, 1.25[
13947	99341	TC093	[09/22/2005, 11/29/2016[	O04	>=1.75	[1, 1.25[
13948	59892	TC093	[09/22/2005, 11/29/2016[	O04	>=1.75	[1, 1.25[
13949	17055	TC093	[09/22/2005, 11/29/2016[	O04	>=1.75	[1, 1.25[
13950	11133	TC091	[09/22/2005, 11/29/2016[	O02	>=1.75	[1.25, 1.5[
13951	15385	TC091	[09/22/2005, 11/29/2016[	O02	>=1.75	[1.25, 1.5[
13952	31222	TC091	[09/22/2005, 11/29/2016[	O02	>=1.75	[1.25, 1.5[
13953	46079	TC091	[09/22/2005, 11/29/2016[	O02	>=1.75	[1.25, 1.5[
13954	52762	TC091	[09/22/2005, 11/29/2016[	O03	>=1.75	[1.25, 1.5[
13955	59702	TC091	[09/22/2005, 11/29/2016[	O03	>=1.75	[1.25, 1.5[

Fig. 24 - Cenário 2 – Excerto da tabela de dados de-identificados - modelo 2

Em suma, ambos os modelos atingiram o objetivo de salvaguarda da identidade dos clientes, reduzindo significativamente os riscos de quebra de privacidade. De notar que, uma vez que o sistema destinatário destes dados de-identificados corresponde a um ambiente de testes, será necessário assegurar que o mesmo representa fielmente a realidade do sistema produtivo, ao mesmo tempo que não coloca em risco a privacidade da informação.

Por se tratar de uma base de dados com um volume significativo de registos (14679) foi possível garantir o agrupamento de diversos *outliers*, mantendo reduzido o número de registos suprimidos.

No que diz respeito à qualidade dos dados finais, apenas o segundo modelo (*K-Anonymization*) permitiu salvaguardar grande parte da diversidade dos dados originais. Não obstante, se necessário, mesmo neste modelo a diversidade poderá ser maximizada com a adição de registos fictícios num novo grupo equivalente, de forma a garantir que o ambiente de testes está adequado à heterogeneidade dos dados reais.

Por fim, e conforme indicado em *[5.2.2 Metodologia Proposta],* a adoção deste modelo de privacidade implicará um trabalho posterior de redistribuição dos dados de--identificados pelas tabelas relacionais de destino, de modo a permitir a sua utilização nos sistemas aplicacionais de testes.

5.3 Cenário 3 - Base de Dados de Trabalhadores Externos

A base de dados em causa inclui informações pessoais de todos os trabalhadores externos que prestam serviços numa Empresa. Para ser considerado apto para o exercício laboral, cada trabalhador tem de demonstrar o cumprimento de um conjunto de requisitos legais e profissionais. Estes dados são armazenados com vista a controlar as autorizações de trabalho, de forma a combater a sinistralidade e trabalho ilegal nas áreas da indústria e construção.

Os referidos dados incluem aptidões médicas, formações de segurança, vistos de residência, idade, etc..

5.3.1 Objetivos

O acesso a estes dados é realizado através de um conjunto de aplicações informáticas integradas entre si. Além do ambiente produtivo, a Empresa em causa necessita também de possuir um *subset* dos dados de trabalhadores externos nos sistemas informáticos de testes (ambientes de desenvolvimento e qualidade).

De notar, no entanto, que não é considerado requisito neste processo de anonimização garantir que os dados sejam representativos da realidade do ambiente produtivo, sendo essencialmente necessário que estes respeitem a mesma natureza, sintaxe e tipologia dos dados originais.

Por outro lado, apenas são necessários cerca de 25% dos registos existentes no sistema produtivo, de forma a evitar uma sobrecarga desnecessária nos sistemas não produtivos (já que os mesmos possuem recursos de *hardware* limitados).

5.3.2 Metodologia Proposta

Após uma análise dos dados existentes no ambiente produtivo foi possível concluir que, na maioria dos casos, não existem ligações diretas entre os diferentes atributos de cada registo, pelo que o tratamento dos dados poderá ser realizado maioritariamente de forma independente sobre cada atributo. Este pressuposto foi posteriormente confirmado pela área de negócio da Empresa.

Assim, foi proposta a aplicação de um conjunto de técnicas de pseudonimização, devidamente adaptadas à tipologia de cada atributo de dados.

Visando mitigar os riscos de ataque identificados no capítulo *4. Falhas na anonimização]*, a proposta preconizou que todos os identificadores ou *quasi-identifiers* fossem tratados usando técnicas baseadas em aleatoriedade ou máscara de caracteres.

De notar que a metodologia proposta é um caso extremo de pseudonimização, já que praticamente todos os dados originais foram substituídos por dados públicos (e.g. nomes próprios), ou por dados tendencialmente aleatórios, tendo por base a envolvente de valores e categorias possíveis para o atributo em causa.

Para garantir a compatibilidade da aplicação informática foram também respeitadas as tipologias, sintaxes e dimensões máximas associadas a cada atributo.

5.3.3 Análise e tratamento prévio dos dados

A base de dados original possuía 25 atributos e 4599 registos, sendo particularmente evidente que muitos destes possuíam dados incompletos: as principais lacunas foram identificadas nos atributos de morada, contactos telefónicos e documentação individual. A empresa confirmou que estes dados estavam efetivamente omissos, como consequência de incompatibilidades verificadas numa importação aplicacional anterior, e que qualquer tratamento dos dados deveria procurar manter a mesma proporcionalidade de registos completos.

Antes de se avançar com o processo de pseudonimização foi necessário realizar uma limpeza dos dados, sendo os casos mais comuns os seguintes:

- Uniformização de valores de categorias:
    - O atributo ESTADOCIVIL estava parametrizado na aplicação como “texto livre”, pelo que existiam inúmeros textos para um conjunto limitado de opções. A título de exemplo, a opção “SOLTEIRO” possuía as seguintes representações: SOL, SOLT, SOLTEI, SOLTE, SOLTEIRO, SOLTEIR0;
    - O atributo LOCALIDADE não estava ligado a uma lista de freguesias e concelhos nacionais, pelo que possuía os mesmos dados (e.g. “V.N.Gaia”) redigidos de inúmeras formas distintas;
    - etc;

- Datas inválidas:
    - Formatação não coerente;
    - DATAENTRAD num futuro longínquo (e.g. 2050);.

Finalizada a limpeza de dados e analisados os diferentes atributos, foi proposta a aplicação das técnicas de pseudonimização constantes da Tabela 21:

Atributo	% Reg.	Técnica Pseudonimização	Dados de referência
ID	100%	<não alterado>	
NOME	100%	*Random Lookup Substitution* (2 nomes próprios + 2 apelidos)	<lista de nomes próprios do IRN [45], + lista dos principais apelidos [46]>
SEXO	100%	<não alterado>	
DATANascim	100%	*Random Date Period Substitution*	<entre min e max do *dataset* inicial>
CodFornecedor	100%	<não alterado>	
DesigFornecedor	100%	*Character Masking*	EMPRESA %CodFornecedor% xxxxxxxx
ACTIVO	100%	<não alterado>	
LOCALTRAB	100%	*Random Lookup Substitution*	<lista de valores do *dataset* inicial>
ESTADOCIVIL	100%	*Random Lookup Substitution*	<lista de valores do *dataset* inicial>
MORADA	14%	*Character Masking*	RUA xxxxxx
CPOSTAL		*Number Substitution*	1234-567
LOCALIDADE		*Random Lookup Substitution*	<lista de valores do *dataset* inicial>
PAIS		*Random Lookup Substitution*	<lista de valores do *dataset* inicial>
TELEFONE	2%	*Number Substitution*	910000000
CATPROFISS	100%	*Random Lookup Substitution*	<lista de valores do *dataset* inicial>
DATAENTRAD	100%	*Random Date Period Substitution*	<entre min e max do *dataset* inicial>
DOCID_TIPO	100%	*Random Lookup Substitution*	<lista de valores do *dataset* inicial>
DOCID_NR	76%	*Random Substitution: Number Generator*	{B} + 9 dígitos *random*
DOCID_VALID	82%	*Random Date Period Substitution*	<entre min e max do *dataset* inicial>

Atributo	% Reg.	Técnica Pseudonimização:	Dados de referência
NUMFISCAL	82%	*Random Substitution: Number Generator + CheckDigit*	{1;2} + 7 dígitos *random* + *checkdigit*
NUMPREVIDENCIA	82%	*Random Substitution: Number Generator + CheckDigit*	{1} + 9 dígitos *random* + *checkdigit*
OBSERVAÇÕES	7%	*Character Masking*	Xxxxxxxxxx
EXAME_MED_DATA	83%	*Random Date Period Substitution*	<entre min e max do *dataset* inicial>
EXAME_MED_RESULT	83%	*Random Lookup Substitution*	{Apto;Apto Condicionalmente}
FORMACAO_SEG_DATA	100%	*Random Date Period Substitution*	<entre min e max do *dataset* inicial>

Tabela 21 - Cenário 3 - Técnicas de pseudonimização propostas

Para substituição dos nomes foram preparadas três listas: nomes próprios masculinos, nomes próprios femininos [45] e apelidos [46]. Todas foram filtradas de modo a incluir apenas os 100 registos mais frequentes em Portugal, numa distribuição uniforme.

Os algoritmos implementados nos atributos NUMFISCAL e NUMPREVIDENCIA permitiram gerar números aleatórios obedecendo à sintaxe normativa de construção de NIF (conforme o Decreto-Lei n.º 14/2013 [47]) e NISS [48], incluindo os respetivos dígitos de controlo.

Sempre que necessário foram ajustadas as funções de "*Random Lookup Substitution*" de modo a garantir a proporcionalidade dos resultados obtidos: a título de exemplo refira-se o atributo "EXAME_MED_RESULT", o qual possui apenas duas opções possíveis ("Apto"; "Apto Condicionalmente"), sendo que "Apto" ocorre em 98.6% dos casos.

Todos os novos registos obedeceram às métricas de máximo número de caracteres, conforme parametrizado na estrutura de atributos da base de dados.

Numa segunda fase foram analisados os resultados obtidos, de modo a validar a dependência semântica entre atributos do mesmo registo. Esta questão foi particularmente evidenciada face à necessidade de coerência entre as diferentes datas associadas a cada indivíduo, conforme patente na Tabela 22:

DATAENTRAD >= DATANascim + 16anos
DOCID_VALID >= DATAENTRAD
DATAENTRAD – 2anos <= EXAME_MED_DATA <= DATAENTRAD + 15dias
DATAENTRAD – 3anos <= FORMACAO_SEG_DATA <= DATAENTRAD

Tabela 22 - Cenário 3 - Regras de Dependência Semântica entre datas do mesmo registo

5.3.4 Resultados

Foram gerados 4599 registos obedecendo às regras apresentadas em 5.3.2., tendo-se posteriormente selecionado (de forma aleatória) 1150 registos (25%), obedecendo à proporcionalidade anteriormente identificada.

A título de exemplo é apresentada na Tabela 23 a comparação entre um registo original e o respetivo correspondente pseudonimizado. Para efeitos de apresentação neste documento os dados originais foram parcialmente alterados, garantindo assim a privacidade do indivíduo.

Atributo	Valor original	Valor pseudonimizado
Id	4003	4003
Nome	MARIA MANUELA SILVA	AUGUSTA VANESSA GONÇALVES BAPTISTA
Sexo	F	F
Datanascim	1982-03-20	1987-01-29
Codfornecedor	61254	61254
Desigfornecedor	ABC Catering Lda	EMPRESA 4003 xxxxxxxxxxx
Activo	N	N
Localtrab	ZONA 10	ZONA 4
Estadocivil	SOLTEIRO	SOLTEIRO
Morada	Lugar da Igreja, 28	RUA 4003 xxxx
Cpostal	5690-276	1234-567
Localidade	MARCO DE CANAVESES	NOGUEIRA DA REGEDOURA
Pais	PORTUGAL	PORTUGAL
Telefone	920392395	910000000
Catprofiss	AJUDANTE DE COZINHA	ASSISTENTE LOGISTICA
Dataentrad	2004-10-20	2013-05-10
Docid_tipo	BI	BI
Docid_nr	11423534	B302994137
Docid_valid	2008-12-13	2015-10-13
Numfiscal	297099249	180713329
Numprevidencia	13280764294	10379277467
Observações	EXAME MÉDICO ANTERIOR 22-10-2002	Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
Exame_med_data	2004-05-29	2011-08-13
Exame_med_result	APTO	APTO
Formacao_seg_data	2004-10-20	2013-03-29

Tabela 23 - Cenário 3 - Pseudonimização de um registo (valor original vs final)

A base de dados resultante do processo de pseudonimização foi apresentada à área de negócio da Empresa, tendo sido confirmada a total adequabilidade e viabilidade para utilização em ambientes de testes (desenvolvimento e qualidade). Mesmo não

constituindo dados reais, todos garantem técnica e funcionalmente os requisitos aplicacionais, estando agora desprovidos de informação confidencial e sensível (dados pessoais de trabalhadores externos). Como tal, foi considerado que o objetivo foi integralmente atingido.

Adicionalmente foi elogiado o facto de os dados manterem um “aspeto realista”, ao contrário dos resultados obtidos com anteriores aplicações de *masking* que optavam, maioritariamente, por substituir os dados sensíveis pelo caractere asterisco (*).

Não tendo sido adotadas técnicas clássicas de anonimização, não é facilmente mensurável o risco de re-identificação destes registos pessoais (tal como indicado em *[3.2 Pseudonimização]*). Não obstante, admitindo que não ocorreram erros na implementação do modelo e que foram eliminadas todas as evidências da operação de pseudonimização, parece seguro adiantar-se que será praticamente impossível inferir qualquer registo original (mesmo que parcialmente), a partir desta base de dados. No entanto, tal não significa que esta nova base de dados possa ser publicamente divulgada: persistem questões de confidencialidade empresarial, uma vez que continuam disponíveis as envolventes dos dados originais (e.g. Categorias Profissionais, Datas de Entrada, Nacionalidades, etc.).

5.4 Cenário 4 - Base de Dados de Colaboradores

Independentemente da dimensão, a maioria das empresas possui uma base de dados de colaboradores onde habitualmente são registados os diversos dados pessoais dos trabalhadores com quem possuem (ou possuíram) vínculos laborais.

Inicialmente sob a alçada dos departamentos de Recursos Humanos, cabe às empresas garantir que qualquer disseminação destes dados pelas restantes áreas da empresa é efetuada sob rigoroso controlo, ao abrigo das utilizações contratualmente previstas.

Caso esse controlo não seja viável de ser implementado, e a partilha dos dados seja efetivamente imprescindível, caberá à empresa filtrar previamente os dados adotando mecanismos que acautelem a privacidade dos indivíduos.

Neste caso específico, a Empresa disponibilizou uma base de dados com informações dos respetivos colaboradores, onde se incluem atributos como o nome, morada, data de nascimento, telefone, nº documento de identificação, entre outros.

5.4.1 Objetivos

Ao contrário dos anteriores cenários, neste caso não foi claramente definido pela Empresa qual o objetivo a dar aos dados referidos. Embora sendo conhecido que parte desta informação é regularmente sincronizada com outras aplicações departamentais (quer em ambientes produtivos, quer de desenvolvimento/qualidade), o desafio lançado visou explorar diferentes técnicas de anonimização, com o objetivo de encontrar opções de alinhamento com a nova regulamentação europeia, procurando sempre que possível maximizar a utilidade e veracidade dos dados.

5.4.2 Metodologia proposta

Face ao repto lançado foram desenvolvidos dois modelos de privacidade tendo em vista a de-identificação dos dados pessoais:

- ➔ De-identificação por “*K-Anonymity*”
- ➔ De-identificação por “*Differential Privacy*”

Foram selecionados dois modelos de de-identificação distintos por se entender que a componente “quantificação de risco” terá um forte papel na decisão final de tratamento dos dados.

No capítulo *[5.4.4 Resultados]* serão avaliados os resultados obtidos e apresentadas as principais conclusões.

5.4.3 Análise e tratamento prévio dos dados

Analisada a estrutura de dados originais foram identificadas as caraterísticas expressas na Tabela 24.

Nº de registos	1808
Nº de atributos	18
Lista de atributos	ID, Nome, NomeConhecido, Título, Email, Email2, EmailGrupo, Sigla, InfoAdicional, Login, Computador, Filial, Ativo, NumFisc, Morada, Telefone, BI e DataNascimento

Tabela 24 - Cenário 4 – Caraterização da base de dados de origem

Face à inexistência dos atributos “localidade” ou “distrito”, e uma vez que a morada completa permitiria uma identificação unívoca do colaborador, foi criado um novo atributo designado por “codpostal”, o qual foi preenchido por desagregação da morada de cada registo (Fig. 26).

Verificou-se também que diversos atributos estavam apenas residualmente preenchidos (<40% registos), pelo que apenas foram considerados e importados para o software ARX os 8 atributos indicados na Fig. 25.

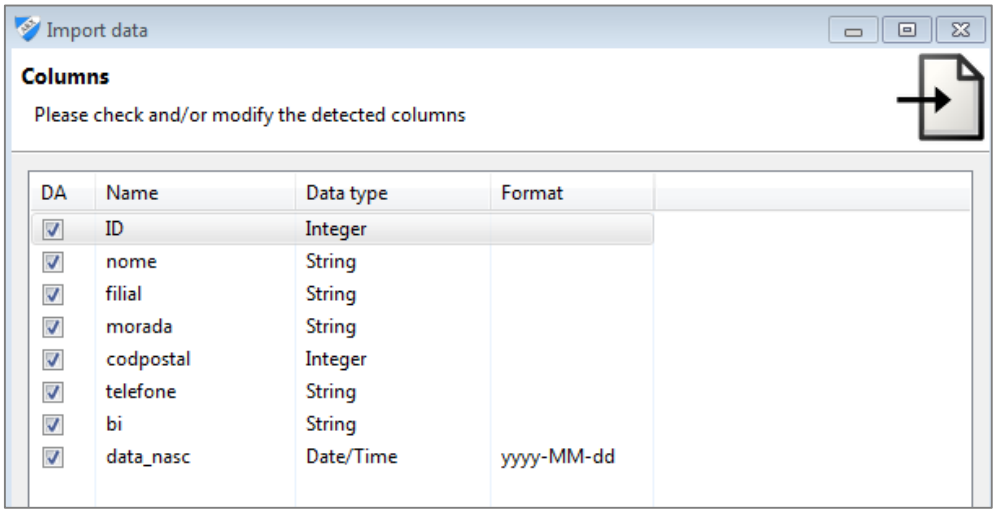


Fig. 25 - Cenário 4 – Importação de dados para o ARX

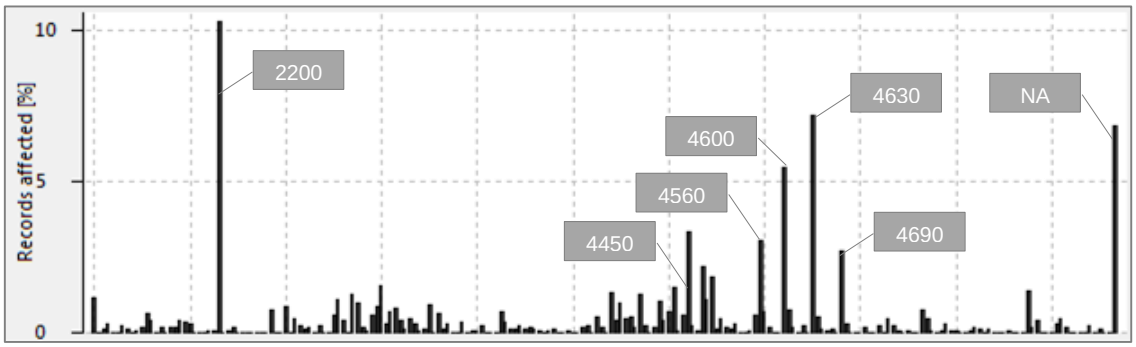


Fig. 26 - Cenário 4 - Histograma de frequência do atributo “codpostal” (dados base)

À semelhança dos cenários analisados anteriormente, o risco inicial é bastante elevado (Fig. 27), tendo-se constatado que alguns dos *quasi-identifiers* permitiam uma identificação quase unívoca; a título de exemplo, o atributo “data_nasc”, de forma isolada, permitiria identificar 74% dos colaboradores.

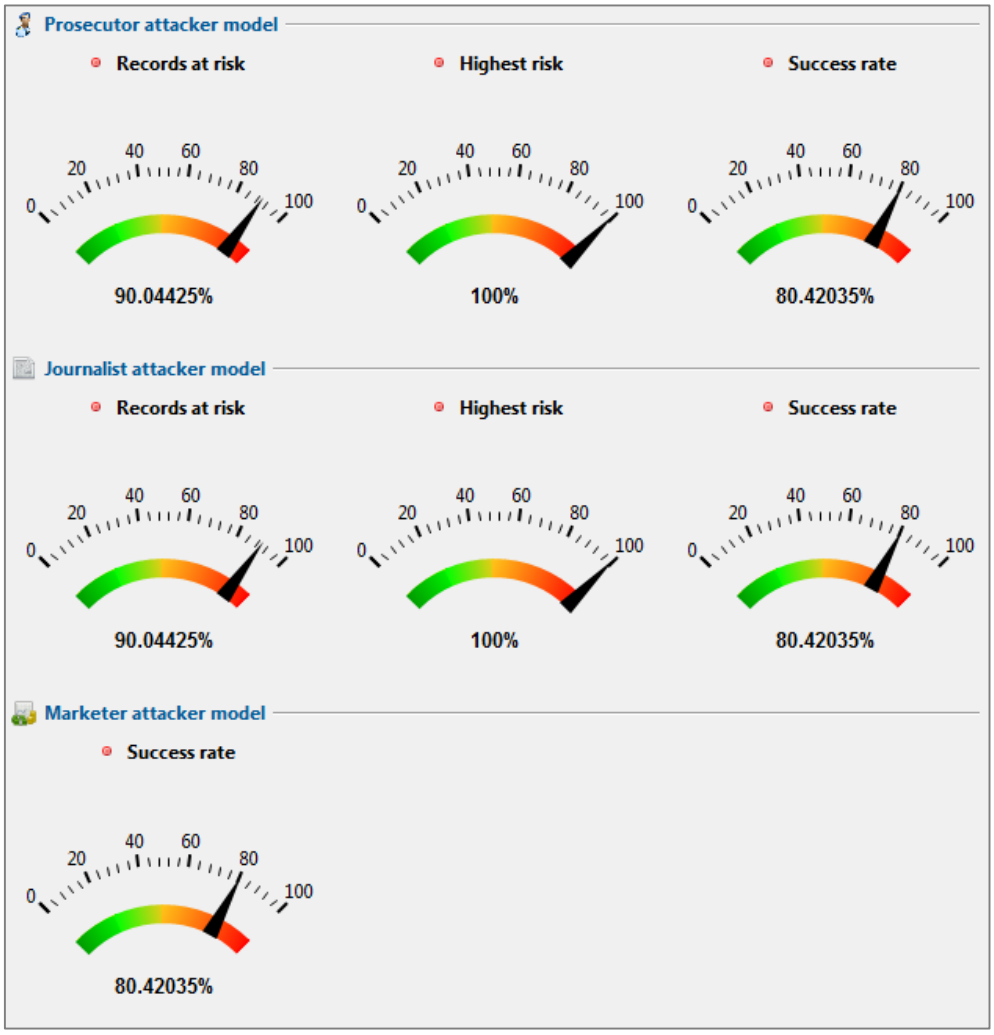


Fig. 27 - Cenário 4 – Riscos de quebra de privacidade (dados base)

Tendo como objetivo a limitação dos riscos de quebra de privacidade, foram inicialmente tipificados os atributos de acordo com a Tabela 25.

Atributo	% Reg	Tipo	Técnica de Generalização	Níveis
ID	100%	*Insensitive*	--	--
nome	100%	Identificador	Máscara de caracteres	0-55
filial	100%	*Quasi-identifier*	N/A	
morada	95%	*Quasi-identifier*	Máscara de caracteres	0-83
codpostal	95%	*Quasi-identifier*	Intervalos de valores	0-6
telefone	47%	*Quasi-identifier*	Máscara de caracteres	0-10
bi	91%	Identificador	--	--
data_nasc	93%	*Quasi-identifier*	Intervalos de datas	0-6

Tabela 25 - Cenário 4 – Tipificação de dados e técnicas de generalização propostas

A título de exemplo, é apresentado na Fig. 28 um excerto da hierarquia de generalização adotada para o atributo “data_nasc”.

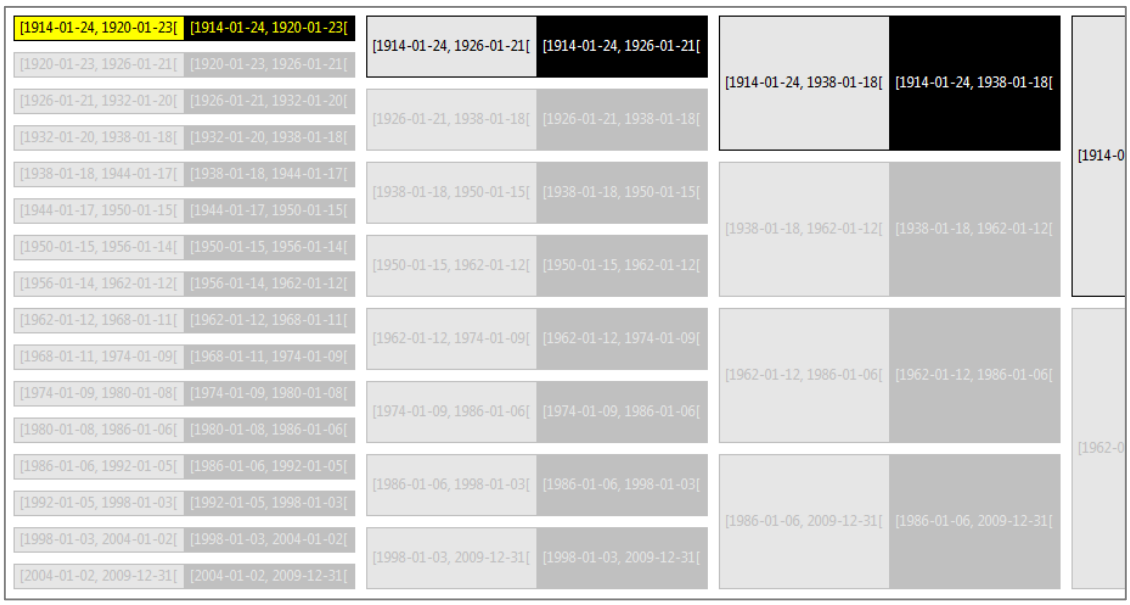


Fig. 28 - Cenário 4 – Hierarquia de Generalização do atributo “data_nasc”

De forma a dar cumprimento à análise exploratória previamente indicada, foram desenvolvidos e aplicados sete modelos de privacidade (Tabela 26), todos eles incidindo sobre o mesmo conjunto de dados base:

Nr	*Privacy Model*	*Supression Limit*	*Quasi-identifiers*
1	*K-Anonymity* com K=2	5%	3: filial, codpostal e data_nasc
2	*K-Anonymity* com K=2	25%	3: filial, codpostal e data_nasc
3	*K-Anonymity* com K=2	50%	6: nome, filial, morada, codpostal, telefone e data_nasc

Nr	*Privacy Model*	*Supression Limit*	*Quasi-identifiers*
4	*K-Anonymity* com K=2	50%	6: nome, filial, morada, codpostal, telefone e data_nasc
5	*K-Anonymity* com K=3	5%	3: filial, codpostal e data_nasc
6	*Differential Privacy* com generalização média	25%	3: filial, codpostal e data_nasc
7	*Differential Privacy* com generalização mínima	Sem limite	3: filial, codpostal e data_nasc

Tabela 26 - Cenário 4 – Parâmetros dos modelos de privacidade

5.4.4 Resultados

De seguida serão apresentados os resultados práticos obtidos com os sete modelos aplicados à base de dados de colaboradores. No final deste capítulo será apresentada uma tabela comparativa e uma avaliação realizada pela Empresa.

- **Cenário 4 – Modelo 1**

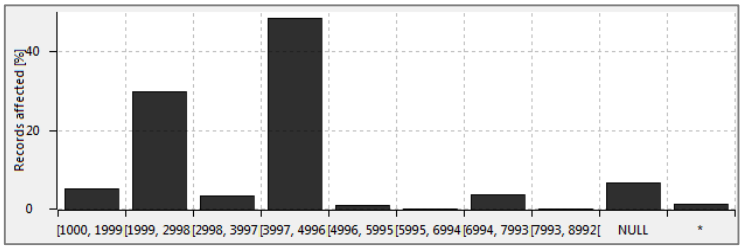
Este primeiro modelo reflete um conjunto de parâmetros conservadores (Tabela 27), servindo também como "grupo de controlo" dos restantes modelos de privacidade.

Cenário 4 – modelo 1 3 *quasi-identifiers*; *K-Anonymity*=2	
% de registos suprimidos	1.4%
Nº de classes Equivalentes	75
Transformação aplicada	[0,1,1]
Generalização do atributo "codpostal"	Nível 1: o código-postal é apresentado como um intervalo de 1000 valores
Generalização do atributo "data_nasc"	Nível 1: a data de nascimento é apresentada como um intervalo de 6 anos
Risco médio de re-identificação	4.15%

Tabela 27 - Cenário 4 – Resultados gerais após de-identificação – modelo 1

Os resultados (apresentados na Fig. 29) validam a viabilidade do modelo, muito embora ocorra perda de informação devido à generalização dos atributos "codpostal" e "data_nasc".

ID	nome	filial	morada	codpostal	ta..	bi	data_nasc
2536	*	Lisboa1	*	[1000, 1999[	*	*	[1926-01-21, 1932-01-20[
2006	*	Lisboa1	*	[1000, 1999[	*	*	[1926-01-21, 1932-01-20[
2011	*	Lisboa1	*	[1000, 1999[	*	*	[1932-01-20, 1938-01-18[
2037	*	Lisboa1	*	[1000, 1999[	*	*	[1932-01-20, 1938-01-18[
2008	*	Lisboa1	*	[1000, 1999[	*	*	[1932-01-20, 1938-01-18[
2411	*	Lisboa1	*	[1000, 1999[	*	*	[1932-01-20, 1938-01-18[
2004	*	Lisboa1	*	[1000, 1999[	*	*	[1932-01-20, 1938-01-18[
2370	*	Lisboa1	*	[1000, 1999[	*	*	[1932-01-20, 1938-01-18[
2005	*	Lisboa1	*	[1000, 1999[	*	*	[1932-01-20, 1938-01-18[
2785	*	Lisboa1	*	[1000, 1999[	*	*	[1938-01-18, 1944-01-17[
2869	*	Lisboa1	*	[1000, 1999[	*	*	[1938-01-18, 1944-01-17[
2850	*	Lisboa1	*	[1000, 1999[	*	*	[1938-01-18, 1944-01-17[
2771	*	Lisboa1	*	[1000, 1999[	*	*	[1944-01-17, 1950-01-15[
2622	*	Lisboa1	*	[1000, 1999[	*	*	[1944-01-17, 1950-01-15[



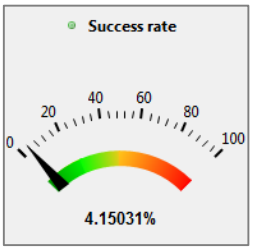


Fig. 29 - Cenário 4 – Excerto de dados de-identificados e Histograma do atributo “codpostal” (modelo 1)

- **Cenário 4 – Modelo 2**

Este segundo modelo é praticamente idêntico ao anterior, sendo apenas aumentado o limite máximo de registos suprimidos, conforme apresentado na Tabela 28.

Cenário 4 – modelo 2 3 *quasi-identifiers*; *K-Anonymity*=2	
% de registos suprimidos	20.2%
Nº de classes Equivalentes	298
Transformação aplicada	[0,0,1]
Generalização do atributo “codpostal”	Nível 0 (generalização mínima), permitindo manter toda a informação deste dado
Generalização do atributo “data_nasc”	Nível 1: a data de nascimento é apresentada como um intervalo de 6 anos
Risco médio de re-identificação	20.59%

Tabela 28 - Cenário 4 – Resultados gerais após de-identificação – modelo 2

Foi assim possível aumentar a qualidade dos dados, nomeadamente pela recuperação de todos os valores do atributo “codpostal” (Fig. 30).

ID	nome	filial	morada	codpostal	...	bi	data_nasc
973	*	Porto	*	4510	*	*	[1980-01-08, 1986-01-06[
872	*	Porto	*	4510	*	*	[1980-01-08, 1986-01-06[
990	*	Porto	*	4510	*	*	[1980-01-08, 1986-01-06[
30	*	Porto	*	4550	*	*	[1938-01-18, 1944-01-17[
44	*	Porto	*	4550	*	*	[1938-01-18, 1944-01-17[
913	*	Porto	*	4550	*	*	[1956-01-14, 1962-01-12[
253	*	Porto	*	4550	*	*	[1956-01-14, 1962-01-12[
931	*	Porto	*	4550	*	*	[1956-01-14, 1962-01-12[
775	*	Porto	*	4550	*	*	[1968-01-11, 1974-01-09[
181	*	Porto	*	4550	*	*	[1968-01-11, 1974-01-09[
132	*	Porto	*	4550	*	*	[1968-01-11, 1974-01-09[
769	*	Porto	*	4550	*	*	[1968-01-11, 1974-01-09[
858	*	Porto	*	4550	*	*	[1968-01-11, 1974-01-09[

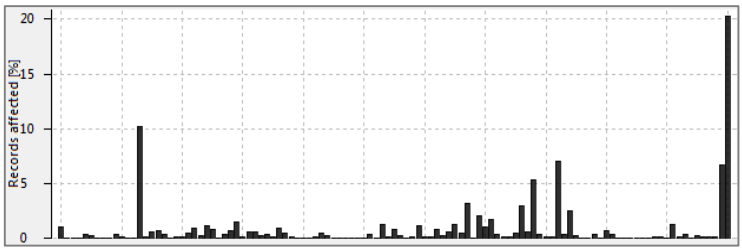


Fig. 30 - Cenário 4 – Excerto de dados de-identificados e Histograma do atributo “codpostal” (modelo 2)

- **Cenário 4 – Modelo 3**

Nesta primeira abordagem com seis *quasi-identifiers* foi forçada a obtenção de resultados mínimos equivalentes (em termos de utilidade dos dados) face ao exercício anterior, garantindo desta forma a manutenção do código postal completo, assim como da data de nascimento no 1º intervalo de generalização (Tabela 29).

Cenário 4 – modelo 3 6 *quasi-identifiers*; *K-Anonymity*=2	
% de registos suprimidos	39.3%
Nº de classes Equivalentes	270
Transformação aplicada	[55,0,79,0,10,1]
Generalização do atributo “codpostal”	Nível 0 (generalização mínima), permitindo manter toda a informação deste dado
Generalização do atributo “data_nasc”	Nível 1: a data de nascimento é apresentada como um intervalo de 6 anos
Risco médio de re-identificação	24.50%

Tabela 29 - Cenário 4 – Resultados gerais após de-identificação – modelo 3

Limitando a supressão de registos a um máximo de 50%, constatou-se ser possível adicionar aos dados de-identificados os quatro primeiros caracteres da morada (e.g.:

"AV.D", "PRAC", "ESTR"). Não obstante, o risco médio de re-identificação passou a atingir os 24.5% (Fig. 31).

ID	no	filial		codpostal	telefone	data_nasc
855	*******	Porto	R.DR*********	4470	**********	[1980-01-08, 1986-01-06[
797	*******	Porto	R.DR*********	4470	**********	[1980-01-08, 1986-01-06[
1168	*******	Porto	RUA *********	4470	**********	[1980-01-08, 1986-01-06[
1184	*******	Porto	RUA *********	4470	**********	[1980-01-08, 1986-01-06[
646	*******	Porto	TRAV*********	4480	**********	[1962-01-12, 1968-01-11[
863	*******	Porto	TRAV*********	4480	**********	[1962-01-12, 1968-01-11[
836	*******	Porto	RUA *********	4480	**********	[1974-01-09, 1980-01-08[
934	*******	Porto	RUA *********	4480	**********	[1974-01-09, 1980-01-08[
406	*******	Porto	RUA *********	4490	**********	[1962-01-12, 1968-01-11[
641	*******	Porto	RUA *********	4490	**********	[1962-01-12, 1968-01-11[
990	*******	Porto	RUA *********	4510	**********	[1980-01-08, 1986-01-06[

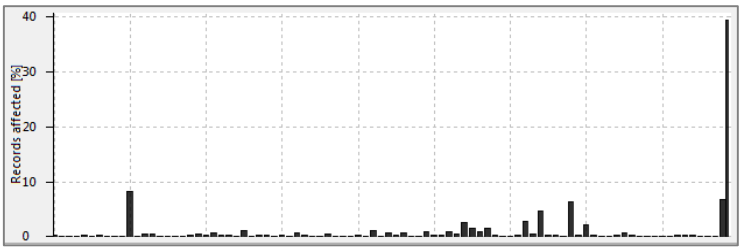


Fig. 31 - Cenário 4 – Excerto de dados de-identificados e Histograma do atributo "codpostal" (modelo 3)

- **Cenário 4 – Modelo 4**

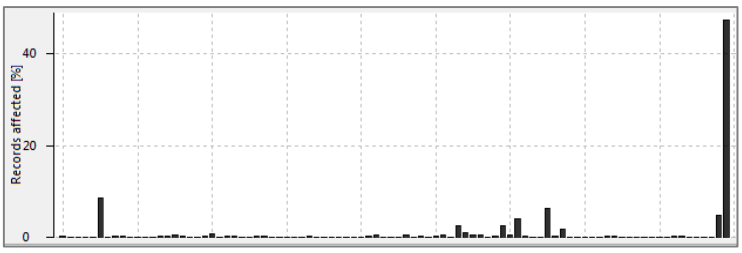
Este modelo tem como prerrogativa a redução da utilidade do atributo "data de nascimento", generalizando-o em intervalos de 24 anos, em alternativa aos 6 anos obtidos no modelo anterior (Tabela 30).

Cenário 4 – modelo 4 6 *quasi-identifiers*; *K-Anonymity*=2	
% de registos suprimidos	47.5%
Nº de classes Equivalentes	279
Transformação aplicada	[53,0,83,0,10,3]
Generalização do atributo "codpostal"	Nível 0 (generalização mínima), permitindo manter toda a informação deste dado
Generalização do atributo "data_nasc"	Nível 3: a data de nascimento é apresentada como um intervalo de 24 anos
Risco médio de re-identificação	29.20%

Tabela 30 - Cenário 4 – Resultados gerais após de-identificação – modelo 4

Desta forma foi possível adicionar aos dados de-identificados os dois primeiros caracteres do nome do colaborador (e.g.: "VI", "MA", "JO" – Fig. 32).

ID	nome	filial		codpostal			data_nasc
2851	VI*********	Lisboa1	*********	2500	***	*	[1938-01-18, 1962-01-12[
2895	VI*********	Lisboa1	*********	2500	***	*	[1938-01-18, 1962-01-12[
6003	MA*********	Lisboa2	*********	NULL	***	*	NULL
6029	MA*********	Lisboa2	*********	NULL	***	*	NULL
128	AB*********	Porto	*********	4560	***	*	[1938-01-18, 1962-01-12[
115	AB*********	Porto	*********	4560	***	*	[1938-01-18, 1962-01-12[
780	AB*********	Porto	*********	4630	***	*	[1962-01-12, 1986-01-06[
725	AB*********	Porto	*********	4630	***	*	[1962-01-12, 1986-01-06[
1172	AD*********	Porto	*********	4000	***	*	[1962-01-12, 1986-01-06[
1204	AD*********	Porto	*********	4000	***	*	[1962-01-12, 1986-01-06[
54	AD*********	Porto	*********	4630	***	*	[1938-01-18, 1962-01-12[



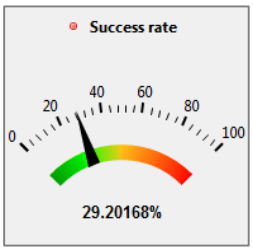


Fig. 32 - Cenário 4 – Excerto de dados de-identificados e Histograma do atributo “codpostal” (modelo 4)

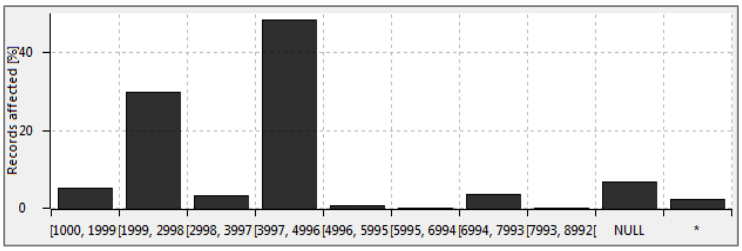
- **Cenário 4 – Modelo 5**

A principal diferença neste modelo (Tabela 31) residiu na adoção de um fator *K-Anonymity*=3, o qual contribuiu para uma descida muito significativa dos riscos de quebra de privacidade, assim como da quantidade de registos em risco. Com a aplicação da transformação [0,1,1] o campo “codpostal” passou a ser generalizado ao 1º nível (intervalos de 1000 valores), perdendo-se alguma utilidade dos dados (Fig. 33).

Cenário 4 – modelo 5 3 *quasi-identifiers*; *K-Anonymity*=3	
% de registos suprimidos	2.3%
Nº de classes Equivalentes	66
Transformação aplicada	[0,1,1]
Generalização do atributo “codpostal”	Nível 1: o código-postal é apresentado como um intervalo de 1000 valores
Generalização do atributo “data_nasc”	Nível 1: a data de nascimento é apresentada como um intervalo de 6 anos
Risco médio de re-identificação	3.69%

Tabela 31 - Cenário 4 – Resultados gerais após de-identificação – modelo 5

ID	nome	filial	morada	codpostal	b..	bi	data_nasc
3017	*	Porto	*	[2998, 3997[	*	*	[1950-01-15, 1956-01-14[
418	*	Porto	*	[2998, 3997[	*	*	[1950-01-15, 1956-01-14[
3018	*	Porto	*	[2998, 3997[	*	*	[1956-01-14, 1962-01-12[
3013	*	Porto	*	[2998, 3997[	*	*	[1956-01-14, 1962-01-12[
921	*	Porto	*	[2998, 3997[	*	*	[1956-01-14, 1962-01-12[
1173	*	Porto	*	[2998, 3997[	*	*	[1962-01-12, 1968-01-11[
1217	*	Porto	*	[2998, 3997[	*	*	[1962-01-12, 1968-01-11[
329	*	Porto	*	[2998, 3997[	*	*	[1962-01-12, 1968-01-11[
522	*	Porto	*	[2998, 3997[	*	*	[1962-01-12, 1968-01-11[
238	*	Porto	*	[2998, 3997[	*	*	[1962-01-12, 1968-01-11[
1125	*	Porto	*	[2998, 3997[	*	*	[1962-01-12, 1968-01-11[
700	*	Porto	*	[2998, 3997[	*	*	[1962-01-12, 1968-01-11[
1000	*	Porto	*	[2998, 3997[	*	*	[1962-01-12, 1968-01-11[
821	*	Porto	*	[2998, 3997[	*	*	[1968-01-11, 1974-01-09[



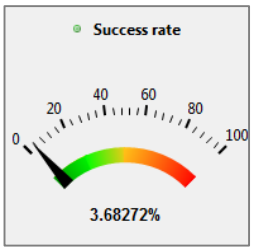


Fig. 33 - Cenário 4 – Excerto de dados de-identificados e Histograma do atributo “codpostal” (modelo 5)

- **Cenário 4 – Modelo 6**

Em contraste com os modelos anteriores, foi adotada neste modelo a técnica “*Differential Privacy*”, com todos os parâmetros padrão (Tabela 32)

Cenário 4 – modelo 6 3 *quasi-identifiers*; Diff Privacy [ε=2, Δ=1e-6, Medium Generalization]	
% de registos suprimidos	18%
Nº de classes Equivalentes	5 (com cerca de 315 registos/classe)
Transformação aplicada	[0,3,3]
Generalização do atributo “codpostal”	Nível 3: o código-postal é apresentado como um intervalo de 4000 valores
Generalização do atributo “data_nasc”	Nível 3: a data de nascimento é apresentada como um intervalo de 24 anos
Risco médio de re-identificação	0.31%

Tabela 32 - Cenário 4 – Resultados gerais após de-identificação – modelo 6

Verificou-se que os índices de risco baixaram para valores quase negligenciáveis, embora tal tenha sido obtido à custa de um elevado nível de generalização (Fig. 34), tornando os resultados parcialmente inúteis.

ID	nome	filial	morada	codpostal	t..	bi	data_nasc
2982	*	Lisboa1	*	[1000, 4996[	*	*	[1962-01-12, 1986-01-06[
2486	*	Lisboa1	*	[1000, 4996[	*	*	[1962-01-12, 1986-01-06[
2459	*	Lisboa1	*	[1000, 4996[	*	*	[1962-01-12, 1986-01-06[
2965	*	Lisboa1	*	[1000, 4996[	*	*	[1962-01-12, 1986-01-06[
2949	*	Lisboa1	*	[1000, 4996[	*	*	[1962-01-12, 1986-01-06[
2402	*	Lisboa1	*	[1000, 4996[	*	*	[1962-01-12, 1986-01-06[
2848	*	Lisboa1	*	[1000, 4996[	*	*	[1962-01-12, 1986-01-06[
3022	*	Porto	*	[1000, 4996[	*	*	[1938-01-18, 1962-01-12[
3017	*	Porto	*	[1000, 4996[	*	*	[1938-01-18, 1962-01-12[
21	*	Porto	*	[1000, 4996[	*	*	[1938-01-18, 1962-01-12[
490	*	Porto	*	[1000, 4996[	*	*	[1938-01-18, 1962-01-12[
60	*	Porto	*	[1000, 4996[	*	*	[1938-01-18, 1962-01-12[

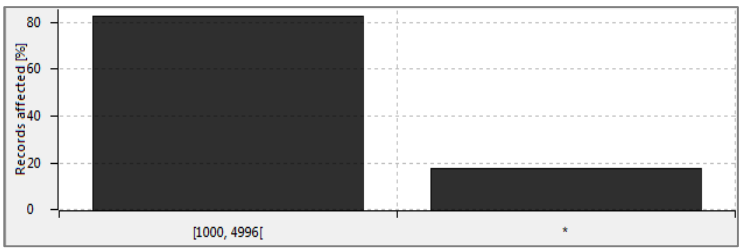


Fig. 34 - Cenário 4 – Excerto de dados de-identificados e Histograma do atributo “codpostal” (modelo 6)

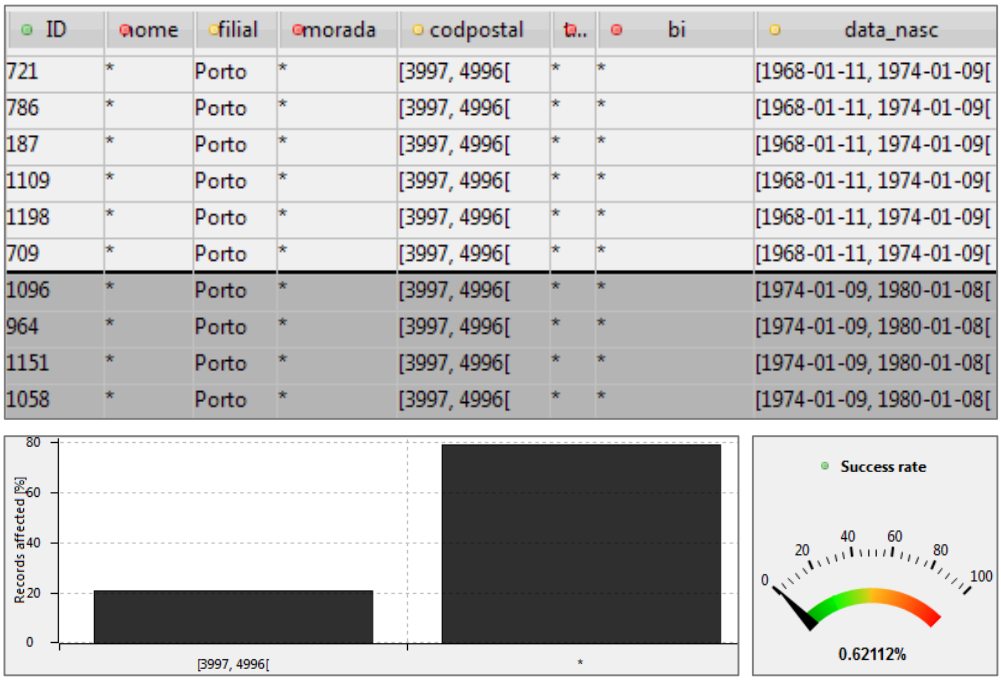
- **Cenário 4 – Modelo 7**

De forma a aumentar a preservação de informação foi realizada uma última iteração, desenvolvendo um modelo “*Differential Privacy*” que force um nível mínimo de generalização (Tabela 33).

Cenário 4 – modelo 7 3 *quasi-identifiers*; Diff Privacy [ε=2, Δ=1e-6, Minimum Generalization]	
% de registos suprimidos	79%
Nº de classes Equivalentes	3
Transformação aplicada	[0,1,1]
Generalização do atributo “codpostal”	Nível 1: o código-postal é apresentado como um intervalo de 1000 valores
Generalização do atributo “data_nasc”	Nível 1: a data de nascimento é apresentada como um intervalo de 6 anos
Risco médio de re-identificação	0.62%

Tabela 33 - Cenário 4 – Resultados gerais após de-identificação – modelo 7

Para ser possível atingir estes resultados, o algoritmo descartou 79% dos dados, agrupando todos os restantes em apenas três classes equivalentes (Fig. 35).



ID	nome	filial	morada	codpostal	ta..	bi	data_nasc
721	*	Porto	*	[3997, 4996[	*	*	[1968-01-11, 1974-01-09[
786	*	Porto	*	[3997, 4996[	*	*	[1968-01-11, 1974-01-09[
187	*	Porto	*	[3997, 4996[	*	*	[1968-01-11, 1974-01-09[
1109	*	Porto	*	[3997, 4996[	*	*	[1968-01-11, 1974-01-09[
1198	*	Porto	*	[3997, 4996[	*	*	[1968-01-11, 1974-01-09[
709	*	Porto	*	[3997, 4996[	*	*	[1968-01-11, 1974-01-09[
1096	*	Porto	*	[3997, 4996[	*	*	[1974-01-09, 1980-01-08[
964	*	Porto	*	[3997, 4996[	*	*	[1974-01-09, 1980-01-08[
1151	*	Porto	*	[3997, 4996[	*	*	[1974-01-09, 1980-01-08[
1058	*	Porto	*	[3997, 4996[	*	*	[1974-01-09, 1980-01-08[

Fig. 35 - Cenário 4 – Excerto de dados de-identificados e Histograma do atributo “codpostal” (modelo 7)

- **Cenário 4 – Comparativo de modelos**

A Tabela 34 apresenta um resumo agregado dos principais resultados obtidos em cada um dos sete modelos de privacidade.

	Modelo Privacidade	Regist Suprim	Precisão CódPostal	Precisão DataNascim	Nome	Morada	Risco Máx.	Risco Méd.
1	K=2 com 3 *quasi-identif.*	1.4%	intervalo de 1000 valores	intervalo de 6 anos	N/A	N/A	50%	4.15%
2	K=2 com 3 *quasi-identif.*	20%	valor original	intervalo de 6 anos	N/A	N/A	50%	20.59%
3	K=2 com 6 *quasi-identif.*	39%	valor original	intervalo de 6 anos	N/A	4 *chars*	50%	24.50%
4	K=2 com 6 *quasi-identif.*	47%	valor original	intervalo de 24 anos	2 *chars*	N/A	50%	29.20%
5	K=3 com 3 *quasi-identif.*	2.4%	intervalo de 1000 valores	intervalo de 6 anos	N/A	N/A	33%	3.70%
6	*Diff.Privacy med. gener.*	18%	intervalo de 4000 valores	intervalo de 24 anos	N/A	N/A	0.5%	0.31%
7	*Diff.Privacy minim gener.*	79%	intervalo de 1000 valores	intervalo de 6 anos	N/A	N/A	0.6%	0.62%

Tabela 34 - Cenário 4 – Resultados gerais após de-identificação – comparativo

Constata-se que o reduzido número de registos (1808) influenciou negativamente os resultados dos diversos modelos de anonimização, já que a inexistência de diversidade nos indivíduos obrigou à considerável adoção de técnicas de generalização ou supressão dos dados.

Com vista a validar o trabalho realizado, assim como aferir a utilidade de aplicação futura das técnicas de anonimização, foram apresentados ao Departamento de Recursos Humanos da Empresa os resultados dos sete modelos.

De uma forma geral foi considerado que todos os modelos cumpriram os objetivos em termos de redução dos riscos de quebra de privacidade. Já no que diz respeito à utilidade e qualidade dos dados, foi designado o modelo 5 (*K-Anonymity* com K=3) como sendo aquele que terá maior potencial para utilização futura, já que combinou simultaneamente as seguintes caraterísticas:

- Foram preservados quase todos os registos da base de dados inicial (a supressão de registos foi inferior a 2.5%);
- A estimativa de risco médio é reduzida (3.7%), quando comparada com as restantes opções;
- A generalização aplicada aos atributos “Código Postal” e “Data de Nascimento” é considerada aceitável (nível 1);

Não foi dada relevância ao facto dos modelos 3 e 4 permitirem manter alguns caracteres dos atributos “nome” e “morada”, já que os mesmos não permitem qualquer utilização prática.

Por fim, foi considerado que alguns dos dados suprimidos durante o processo de de-identificação seriam fundamentais em determinados sistemas externos ao Departamento, pelo que deverão ser promovidas as necessárias atualizações procedimentais e contratuais, com vista a permitir a partilha e sincronização dessa informação com outros departamentos da Empresa, incluindo:

- Assegurar o consentimento informado dos titulares, para os fins pretendidos;
- Dar formação e garantir compromissos de sigilo dos colaboradores;
- Implementar mecanismos de logging, autenticação e autorização;
- Adaptar os sistemas para uma política “*Privacy by Design*”;
- Implementar PIA, assim como avaliação dos riscos de quebra de privacidade e definição de medidas mitigadoras, promovendo frequentes auditorias.

6. Conclusões e Trabalhos Futuros

A privacidade é um direito fundamental reconhecido na Declaração Universal dos Direitos do Homem, sendo "*indispensável para a nossa capacidade de controlar a relação com o mundo*" [49]. Não obstante, ao longo das últimas décadas, múltiplas têm sido as interpretações realizadas pelos diferentes Estados-Membros da UE no que concerne à respetiva aplicação prática, e em especial no que diz respeito aos dados pessoais. O novo Regulamento Geral de Proteção de Dados visa harmonizar o nível de proteção dos direitos e liberdades, impondo regras às empresas de todos os Estados-Membros.

Pretendeu-se com este trabalho explorar algumas das técnicas utilizadas em processos de anonimização, de-identificação e pseudonimização, frequentemente utilizadas no meio empresarial e académico, demonstrando a importância da sua aplicação sobre bases de dados empresariais, visando a proteção de dados pessoais e mitigação dos riscos de quebra de privacidade.

Foram desenvolvidos e aplicados modelos de de-identificação, anonimização e pseudonimização sobre quatro casos práticos de bases de dados reais, utilizando técnicas de *K-Anonymity, L-Diversity, Differential Privacy* e *Data Masking.*

Como ponto de partida, foram avaliados os riscos de re-identificação das bases de dados mantendo todos os *quasi-identifiers* (isto é, suprimindo apenas os atributos que identificavam univocamente cada um dos registos). Em todos os casos foi verificado que esses riscos iniciais eram extremamente elevados, bastando a combinação de dois atributos para potenciar a re-identificação da quase totalidade dos registos.

Constatou-se que a aplicação de técnicas de anonimização a bases de dados relacionais implica um significativo trabalho prévio de conversão das tabelas, uma vez que os algoritmos foram essencialmente concebidos para lidar com tuplos. Esta conversão poderá ter de ser repetida após concluída a anonimização, de modo a recriar a estrutura relacional das bases de dados originais.

Verificou-se que uma parte dos algoritmos de anonimização analisados tende a falhar quando as bases de dados são esparsas, uma vez que não estão otimizados para lidar com dezenas ou centenas de atributos que constituem, na prática, *quasi--identifiers*. Nestas situações será muito difícil garantir simultaneamente a veracidade e a de-identificação de alguns atributos, pelo que se sugere (quando possível) "quebrar" a ligação com os dados de origem, aplicando técnicas de pseudonimização com aleatoriedade.

Após aplicação dos modelos de privacidade foi possível demonstrar que o risco de re-identificação diminuiu substancialmente, chegando a atingir valores ínfimos quando aplicada a técnica de "*Differential Privacy".* Não obstante, independentemente da aparente eficácia do processo de anonimização, será sempre necessário admitir que persistem riscos de re-identificação, por mais residuais que possam parecer. De notar que os modelos de quantificação de risco apresentados neste documento têm como premissa que um potencial atacante não possui qualquer informação adicional sobre os indivíduos. Sempre que tal não se verifique, o risco de re-identificação será tendencialmente superior, e no limite até poderá comprometer a possibilidade de divulgação de quaisquer dados pessoais, conforme exemplificado nos diversos ataques apresentados no capítulo "4. Falhas na anonimização".

Por outro lado, o risco não deverá ser visto exclusivamente como um bloqueio. Os processos de anonimização são fundamentais para viabilizar a disponibilização de determinadas informações para as empresas e para a sociedade, pelo que, em última análise, trabalhar com anonimização de dados pessoais significa gerir risco (e conformidade com as normas e leis em vigor). Caberá sempre às empresas a responsabilidade de reavaliar regularmente os riscos associados aos processos de anonimização, mesmo que estes tenham já ocorrido no passado.

Com o desenvolvimento deste trabalho tornou-se evidente que o desafio não passa apenas pela anonimização ou de-identificação dos dados pessoais, mas também por conseguir assegurar a veracidade (os resultados serem representativos da mesma realidade dos dados originais), a qualidade (minimizar a generalização dos atributos) e utilidade (minimizar a supressão de registos) dos mesmos. Para tal, considera-se fundamental o envolvimento ativo dos destinatários finais da informação, visando a adequada calibração dos modelos de privacidade.

É também de realçar que, mesmo de-identificados ou até eventualmente anonimizados, os dados não deixam de conter informação potencialmente confidencial

da empresa, pelo que continua a ser fundamental precaver obrigações contratuais de sigilo profissional.

Em muitas situações a de-identificação implicará a destruição da qualidade dos dados, pelo que as empresas terão que garantir a conformidade com a nova regulamentação europeia, caso pretendam preservar ou distribuir essa informação. Tal deverá originar ajustes ao nível do consentimento informado para recolha dos dados, formação e sigilo dos trabalhadores, implementação de sistemas seguros e que sigam as melhores práticas de "*Privacy by Design*", assim como promover as respetivas avaliações de risco.

Complementarmente, e de forma a acautelar a conformidade com o GDPR, as empresas poderão também optar por implementar produtos e serviços previamente certificados (e.g. com o selo EuroPriSe), assim como realizar periodicamente *Privacy Impact Assessments (PIA)* de forma a auditar os diversos processos organizacionais.

Como trabalho futuro foi identificada a necessidade de desenvolvimento de um algoritmo que permita quantificar o risco de re-identificação associado a bases de dados pseudonimizadas (e.g. via processos de *data masking*), sugerindo-se como ponto de partida o trabalho desenvolvido em [17] [18] [19].

Por fim, considera-se também importante complementar este trabalho, através da demonstração da aplicação prática destas técnicas de anonimização a bases de dados de organismos públicos, de forma a validar as conclusões obtidas.

Referências bibliográficas

[1] "REGULAMENTO (UE) 2016/679 do Parlamento Europeu e do Conselho - Proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados," 27 de abril de 2016.

[2] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin, "A Systematic Review of Re-Identification Attacks on Health Data," *PLoS ONE 6(12)*, vol. e28071, 2001.

[3] Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Review*, vol. 57, p. 1701, 2010.

[4] Austrian Federal Chancellery, "Federal Act concerning the Protection of Personal Data (DSG 2000)," vol. ERV_1999_1_165, Jan. 2017.

[5] "DIRETIVA (UE) 95/46/CE do Parlamento Europeu e do Conselho - Proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados," 24 de outubro de 1995.

[6] Agência dos Direitos Fundamentais da União Europeia, Conselho da Europa, "Manual da Legislacao Europeia sobre Protecao de Dados," Luxemburgo, 2014.

[7] Comissão Nacional de Proteção de Dados, "10 medidas para preparar a aplicação do Regulamento Europeu de Proteção de Dados," 2017.

[8] Simson L. Garfinkel, "NIST.IR.8053 - De-Identification of Personal Information," 2015.

[9] Gregory S. Nelson, "Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification," ThotWave Technologies, 2015.

[10] Article 29 Data Protection Working Party, "Opinion 05/2014 on Anonymisation Techniques," 2014.

[11] "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule," HHS - Office for Civil Rights (OCR), 2012.

[12] Luk Arbuckle Khaled El Emam, *Anonymizing Health Data: Case Studies and Methods to Get You Started*, 1st ed.: O'Reilly Media, 2013.

[13] S.Y. Esayas, "The role of anonymisation and pseudonymisation under the EU data privacy rules: beyond the "all or nothing" approach," *European Journal of Law and Technology*, vol. 6, Nr.2, 2015.

[14] Comissão Nacional de Proteção de Dados, "Deliberação nº 1704/2015 - Aplicável aos tratamentos de dados pessoais efetuados no âmbito da Investigação Clínica," Lisboa, 22 de outubro de 2015.

[15] Rathindra Sarathy and Krish Muralidhar, "A Common Index Of Similarity For Numerical Data Masking Techniques," *United Nations Statistical Commission And Economic Commission For Europe Conference Of European Statisticians*, vol. Risk/benefit analysis and new directions for statistical disclosure limitation, 2009.

[16] A. C. Singh, F. Yu, and G.H. Dunteman, "Massc: A new data mask for limiting statistical information loss and disclosure," *Work Session on Statistical Data Confidentiality*, vol. Monographs in Official Statistics, 2004.

[17] Eugene W. Myers, "An O(ND) Difference Algorithm and Its Variations," *Department of Computer Science, University of Arizona, Tucson, AZ 85721, U.S.A*.

[18] Esko Ukkonen, "Algorithms for Approximate String Matching," *Information and Control*, vol. 64, pp. Issues 1-3, 1985.

[19] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C. Tappert, "A Survey of Binary Similarity and Distance Measures," *Systemics, Cybernetics and Informatics*, vol. 8 - Nº 1, 2010.

[20] Adam Smith, *Pinning Down "Privacy" in Statistical Databases*, Crypto Tutorial ed. USA: Penn State Computer Science & Engineering Department, 2012.

[21] Cynthia Dwork, "Differential Privacy: A Survey of Results," in *5th International Conference in TAMC (Theory and Applications of Models of Computation)*, 2008.

[22] Balaji Raghunathan, *The Complete Book of Data Anonymization - From Planning to Implementation*.: Infosys Press, 2013.

[23] Arvind Narayanan and Edward W. Felten, "No silver bullet: De-identification still doesn't work," Princeton University - Department of Computer Science, 2014.

[24] Yanni Lagos, "Taking the personal out of data: Making sense of de-identification," Ind. L. Rev. 2014.

[25] Friedewald M., Hansen M., Obersteller H., Rost M. Bieker F., "A Process for Data Protection Impact Assessment Under the European General Data Protection Regulation," *Privacy Technologies and Policy*, vol. 9857, no. Lecture Notes in Computer Science, 2016.

[26] Nora Cuppens-Boulahia, Frédéric Cuppens, Noémie Jess, Françoise Dupont, et al. Maxime Bergeat, "A French Anonymization Experiment with Health Data," in *Privacy in Statistical Databases*, Eivissa, Spain, 2014.

[27] M Barbaro and Jr T. Zeller, "A Face is Exposed for AOL Searcher," *New York Times*, vol. 4417749, Aug 2006.

[28] Spiliopoulou M., Baeza-Yates R. Poblete B., "Website Privacy Preservation for Query Log Publishing," in *Privacy, Security, and Trust in KDD.*, 2008.

[29] Latanya Sweeney, "Matching Known Patients to Health Records in Washington State Data," Harvard University. Data Privacy Lab, 2013.

[30] Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, and Alex "Sandy" Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, pp. 536-539, jan 2015.

[31] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel, "Unique in the Crowd: The privacy bounds," *Scientific Reports 3*, 2013.

[32] Arvind Narayanan and Vitaly Shmatikov, "Robust De-anonymization of Large Sparse Datasets," *In Security and Privacy 2008*.

[33] Latanya Sweeney, "Simple Demographics Often Identify People Uniquely," Carnegie Mellon University, 2000.

[34] Daniel C. Barth-Jones, "The "Re-identification" of Governor William Weld's Medical Information: A Critical Re-examination of Health Data Identification Risks and Privacy Protections, Then and Now," Columbia University - Department of Epidemiology, 2012.

[35] Amy L. McGuire, David Golan, Eran Halperin and Yaniv Erlich Melissa Gymrek, "Identifying Personal Genomes by Surname Inference," *Science*, vol. 339, pp. 321-324, jan 2013.

[36] John Bohannon, "Genealogy Databases Enable Naming of Anonymous DNA Donors," in *Science, Vol. 339, No. 6117*, 2013.

[37] Vijay Pandurangan. (2014, junho) On Taxis and Rainbows - Lessons from NYC's improperly anonymized taxi logs. [Online]. https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1

[38] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith, "Composition attacks and auxiliary information in data privacy," in *KDD '08 Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, Las Vegas, USA, 2008.

[39] Andrew Ruddick and Jeff Yan, "Acceleration Attacks on PBKDF2. Or, what is inside the black-box of oclHashcat?," *WOOT16*, 2016.

[40] Jagan Athreya Waleed Ahmed, "Data Masking Best Practice," ORACLE Corporation, 2013.

[41] Rogério Reis, *Criptografia não Digital - Apontamentos da cadeira de Criptografia do Mestrado de Segurança Informática*. Porto: FCUP, 2014.

[42] AT- Autoridade Tributária e Aduaneira, Manual de Operações - Início de Atividade, junho 2016.

[43] Himanshu Taneja Kapil and Ashutosh Kumar Singh, "Preserving Privacy of Patients based on Re-identification Risk," in *4th International Conference on Eco-friendly Computing and Communication Systems, ICECCS*, 2015.

[44] Privacy Analytics, "Don't Blur the Lines between Data Masking and Real De-identification," White Paper 2017.

[45] Instituto de Registos e Notariado - Ministério da Justiça. (2017, mar) Vocábulos admitidos e não admitidos como nomes próprios. [Online]. http://www.irn.mj.pt/sections/irn/a_registral/registos-centrais/docs-da-nacionalidade/vocabulos-admitidos-e/downloadFile/file/Lista_de_nomes2017-03-31.pdf

[46] Carolina Reis e Ana Serra, "Os 100 apelidos mais comuns em Portugal. No reino dos Silvas, Santos e Pereiras," *Expresso*, outubro 2015.

[47] "Decreto-Lei n.º 14/2013," *Diário da República, 1.ª série, N.º19*, janeiro 2013.

[48] Carlos Galhano. (2017, junho) Check-digit do NISS (Número de Identificação na Segurança Social). [Online]. http://www.galhano.com/blog/wp-content/2007/06/NISS.pdf

[49] Bruce Schneier, *Data and Goliath*.: W.W.Norton & Company, 2015.